PB87-206496

Protocol for Applying and Validating the CMB (Chemical Mass Balance) Model

(U.S.) Environmental Protection Agency
Research Triangle Park, NC

May 87

U.S. Department of Commerce
National Technical Information Service
NTIS

PB87-206496

United States
Environmental Protection
Agency

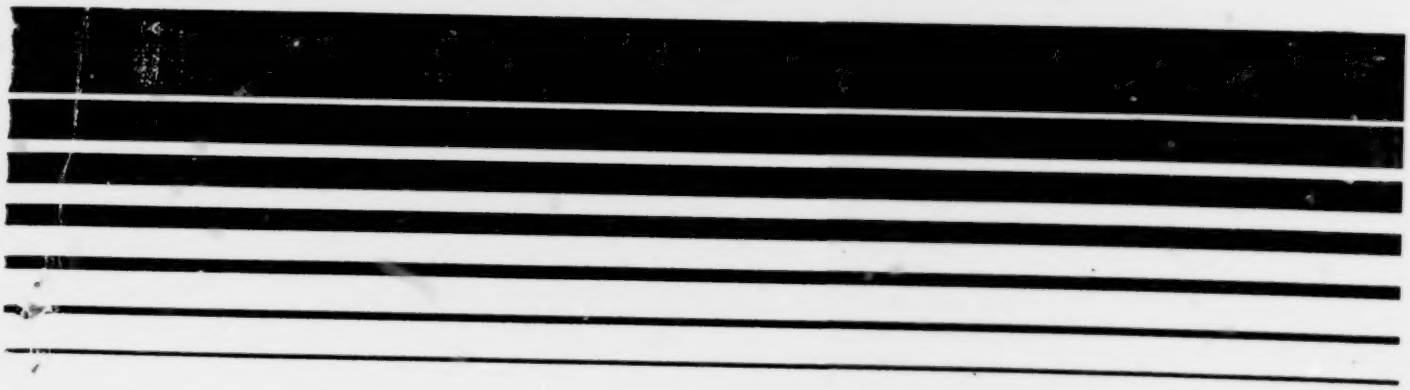
Office of Air Quality
Planning and Standards
Research Triangle Park NC 27711

EPA-450/4-87-010
May 1987

Air

EPA

Protocol For Applying And Validating The CMB Model



REPRODUCED BY
U.S. DEPARTMENT OF COMMERCE
NATIONAL TECHNICAL
INFORMATION SERVICE
SPRINGFIELD, VA 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
		PB87 206496/AS

4. TITLE AND SUBTITLE
Protocol for Applying and Validating the CMB Model

5. REPORT DATE
May 1987

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
Thompson G. Pace and Dr. John G. Watson

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Air Management Technology Branch
Monitoring and Data Analysis Division
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

Desert Research Institute
Reno, NV 89506

10. PROGRAM ELEMENT NO.

11. CONTRACT/GRANT NO.
CX-813087-01-1

12. SPONSORING AGENCY NAME AND ADDRESS
U.S. Environmental Protection Agency
OAQPS, MDAD, MD-14
Research Triangle Park, NC 27711

13. TYPE OF REPORT AND PERIOD COVERED

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT
This protocol is intended to supplement the User's Manual for the CMB model by providing technical guidance on: (a) the selection of model input data, (b) determination of validity and uncertainty of a specific model application, and (c) reduction of the uncertainty associated with the results of a specific application. The objective of a CMB application is to use information about the chemical composition of sources in an airshed (source profiles) along with data on the chemical composition of the ambient air to estimate the source contributions which would best "explain" the chemical properties of measured ambient data (species).

The guidance provided by this document consists of a seven-step process:

1. assessing the general applicability of the CMB model to the situation under study;
2. configuring the model with appropriate sources, source profiles, and chemical species concentrations at receptor sites;
3. examining model statistics and diagnostics;
4. determining agreement with model assumptions;
5. identifying problems, changing the model configuration and rerunning;
6. testing the consistency and stability of model results; and
7. evaluating the validity of model results.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Receptor Models Chemical Mass Balance Source Apportionment Least Squares Multiple Linear Regression		

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
	Unlimited	74
	20. SECURITY CLASS (This page)	22. PRICE
	Unlimited	PC $13.95

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

EPA-450/4-87-010

PROTOCOL FOR APPLYING AND VALIDATING THE CMB MODEL

By

Air Management Technology Branch
Monitoring And Data Analysis Division
And
Desert Research Institute
Reno, Nevada 89506

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office Of Air And Radiation
Office Of Air Quality Planning And Standards
Research Triangle Park NC 27711
May 1987

This report has been reviewed by the Office Of Air Quality Planning And Standards, U.S. Environmental Protection Agency, and approved for publication as received from the contractor. Approval does not signify that the contents necessarily reflect the views and policies of the Agency, neither does the mention of trade names or commercial products constitute endorsement or recommendation for use.

EPA-450/4-87-010

TABLE OF CONTENTS

		Page
List of Figures		v
List of Tables		vi
Executive Summary		vii
1.0	Introduction	1
2.0	Planning the CMB Analysis	3
2.1	Other Useful Receptor Methods	3
2.2	Levels of CMB Analysis	3
3.0	Applications and Validation Protocol	5
3.1	Determining Applicability of the CMB	6
3.2	CMB Model Setup and Initial Run	7
	3.2.1 Ambient Species and Their Uncertainty	7
	3.2.2 Source Identification - The Fitting Sources	9
	3.2.3 Source Profile Representativeness and Uncertainty	10
	3.2.4 Initial Run of the CMB Model	11
3.3	Outputs, Statistics and Diagnostics - Definitions and Interpretation	11
	3.3.1 Percent Total Mass Accounted For	12
	3.3.2 R-Square and Chi-Square -- Goodness of Fit	12
	3.3.3 t-Statistic	16
	3.3.4 Uncertainty/Similarity Clusters	16
	3.3.5 Ratios and Residuals of Fitting Species	17
	3.3.6 Source Species Contributions	18
3.4	Deviations from Model Assumptions	18

Page

3.5 Identifying and Correcting Problems by Changing the Model Inputs..... 19

3.5.1 Correcting the Ambient Data - Gross Errors..... 19

3.5.2 Correcting Source Profiles - Gross Errors..... 20

3.5.3 Changing the Source List..... 20

3.5.3.1 Missing Sources..... 20

3.5.3.2 Noncontributing Sources..... 21

3.5.4 Improving Source Profiles--Uncertainty/Similarity..... 22

3.5.5 Problem Identification and Correction Strategy..... 23

3.6 Consistency/Stability of the Model Results..... 25

3.6.1 Source Profile Sensitivity..... 25

3.6.2 Receptor Concentration Sensitivity..... 26

3.6.3 Fitting Species Sensitivity..... 26

3.7 Evaluating Results of the CMB Analyses..... 26

4.0 Acknowledgments..... 27

5.0 References..... 28

Appendix A - Determining Source Profiles and Their Uncertainties..... A-1

Appendix B - Examples and Quick Guide for Identifying Problems and Corrective Actions..... B-1

Appendix C - CMB Model Assumptions..... C-1

LIST OF FIGURES

Number		Page
A	CMB Application and Validation Flowchart.....	ix
B	Flowchart for Problem Identification and Correction..	x
1	Flowchart for Problem Identification and Correction..	24

LIST OF TABLES

Number		Page
A	Summary of Model Displays (Outputs, Statistics, Diagnostics and Status)	xii
2.1	Useful Receptor Methods	4
3.1	Summary of Model Displays (Outputs, Statistics, Diagnostics and Status)	13

EXECUTIVE SUMMARY

The objectives of this applications and validation protocol are to provide technical guidance on:

- The selection of source categories to be apportioned by the model and the selection of model input data.
- Methods to determine the validity and uncertainty of a specific model application.
- Methods to improve the quality and reduce the uncertainty of a specific model application.

The objective of a CMB application is to use information about the chemical composition of sources in an airshed (source profiles) along with data on the chemical composition of the ambient air to estimate the source contributions which would best "explain" the chemical properties of measured ambient data (species). These source profiles and ambient species data are the primary model inputs. Effective variance weighted least squares, which considers the uncertainties in both source and ambient data is used to derive the source contribution estimates.

The CMB consists of a set of linear equations which express the ambient concentrations of chemical species as the sum of products of source compositions and source contributions. These equations are a direct consequence of the law of conservation of mass. This set of equations is overdetermined because the number of measured chemical species generally exceeds the number of contributing source types. The source contributions are generally the unknowns in these equations. However, a unique solution cannot be found for this set of equations because measurement uncertainty precludes determination of exact values for source and receptor data. When these uncertainties are estimated for both source and receptor measurements, an additional physical constraint is applied which will yield a unique solution. This constraint is the minimization of the difference between calculated and measured receptor concentrations by using a weighting scheme, effective variance. The weighting has a physical significance in that it is derived from the measurement uncertainties of both source and receptor chemical species. Though the CMB solution is identical to some statistical inference methods, it is not entirely dependent on statistical principles. The basic model equations which represent the source receptor relationship, the effective variance weighting, and the error propogation are all based on physical principles which are then incorporated into a statistical estimation procedure.

The guidance provided by this document consists of a seven-step process which is described in Sections 3.1 through 3.7 of this protocol. This guidance is necessary because there is no single goodness-of-fit parameter which can assess the models reliability. Application of the CMB is ideally preceded by other receptor methods or dispersion models.

This is described in Section 2.0. Also, the CMB results must be compared with other receptor and dispersion model results and any differences must be reconciled. Finally, a control strategy must be developed. The model reconciliation and control strategy development processes are described in other protocols. The CMB must have a strong physical basis in order for its results to be meaningful. The seven steps listed below are summarized in Figures A and B of this summary.

(1) assessing the general applicability of the CMB model to the situation under study;

(2) configuring the model with appropriate sources, source profiles, and chemical species concentrations at receptor sites;

(3) examining model statistics and diagnostics;

(4) determining agreement with model assumptions;

(5) identifying problems, changing the model configuration and rerunning;

(6) testing the consistency and stability of model results; and

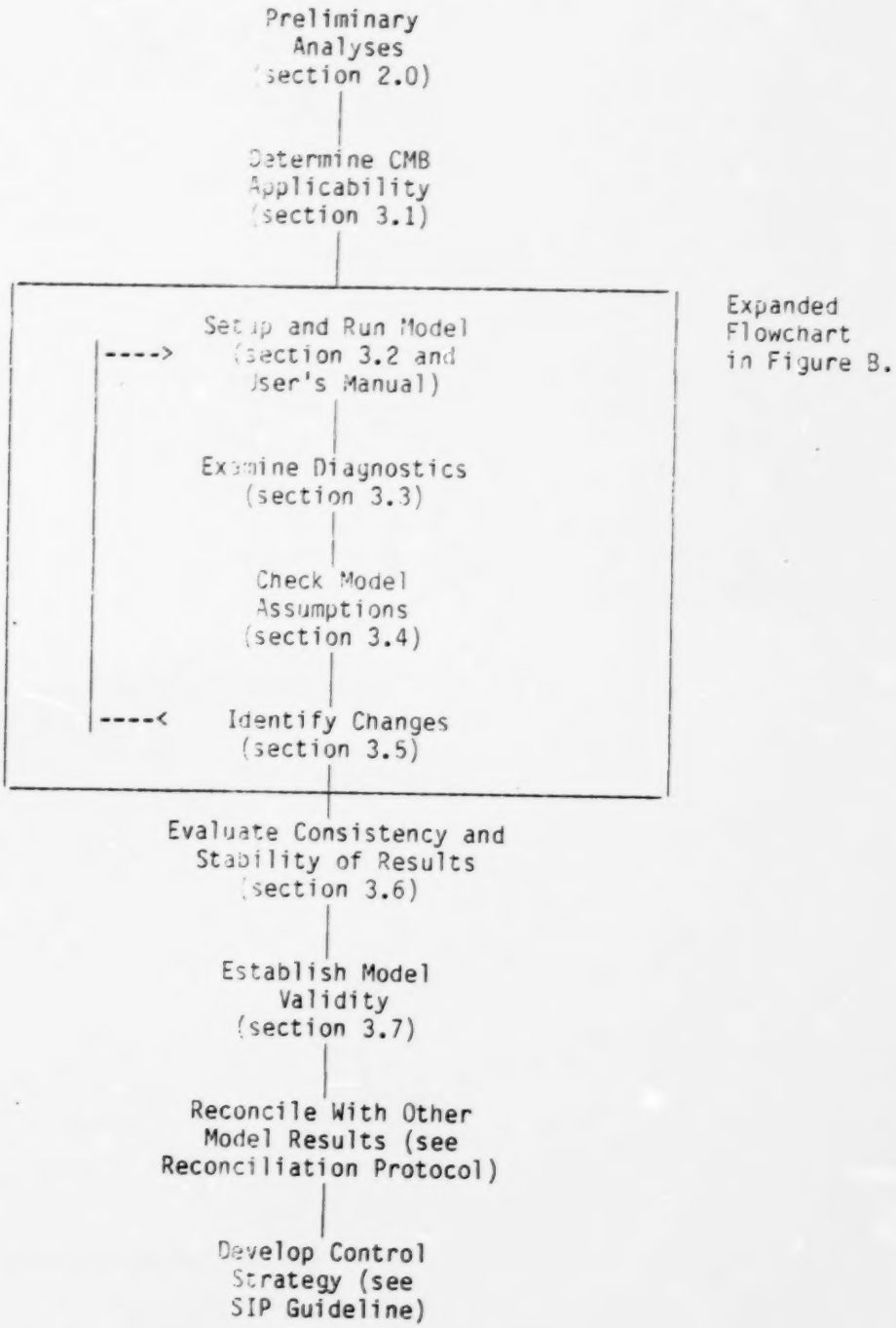
(7) evaluating the validity of model results.

The applicability of the CMB model must be determined for each application. This determination concerns such issues as sample quantity, quality and appropriateness for chemical analysis. Also, the model can only estimate the contributions (to ambient mass concentration) of source categories which are chemically distinguishable (different) from each other. Their chemical characteristics must be obtainable.

The model provides three primary outputs. These are: (1) contribution estimates (to ambient concentrations) of the sources or source categories which are included in the fit (SCE); (2) the standard errors of these source contribution estimates (STDERR); and (3) the species concentrations calculated from the fit (CALC). There are three statistics which can be used to evaluate how well the model's calculated species concentrations "fit" or match the ambient measurements for these species. These are: (1) the percent of total mass explained by the fit (%MASS); (2) R-SQUARE; and (3) CHI-SQUARE. Generally, it is desirable to obtain a good "fit" of the data and obtain SCE's which have STDERR that are low relative to the size of the SCE.

Four diagnostics are available to help identify the data which are responsible for a poor fit so that improved data might be obtained or included to improve the fit. These are: (1) the uncertainty/similarity clusters (U/S Clusters); (2) the ratio of calculated to measured species concentrations (RATIO C/M); the ratio of the residual (calculated-measured) to the uncertainty of this difference (RATIO R/U); and (4) the portion of a calculated species concentration that is attributed by the model to each source (SSCONT). Table A summarizes the information available on the model displays.

Figure A. CMB Application and Validation Flowchart


Preliminary Analyses (section 2.0)
Determine CMB Applicability (section 3.1)
Setup and Run Model (section 3.2 and User's Manual)
Examine Diagnostics (section 3.3)
Check Model Assumptions (section 3.4)
Identify Changes (section 3.5)
Expanded Flowchart in Figure B.
Evaluate Consistency and Stability of Results (section 3.6)
Establish Model Validity (section 3.7)
Reconcile With Other Model Results (see Reconciliation Protocol)
Develop Control Strategy (see SIP Guideline)

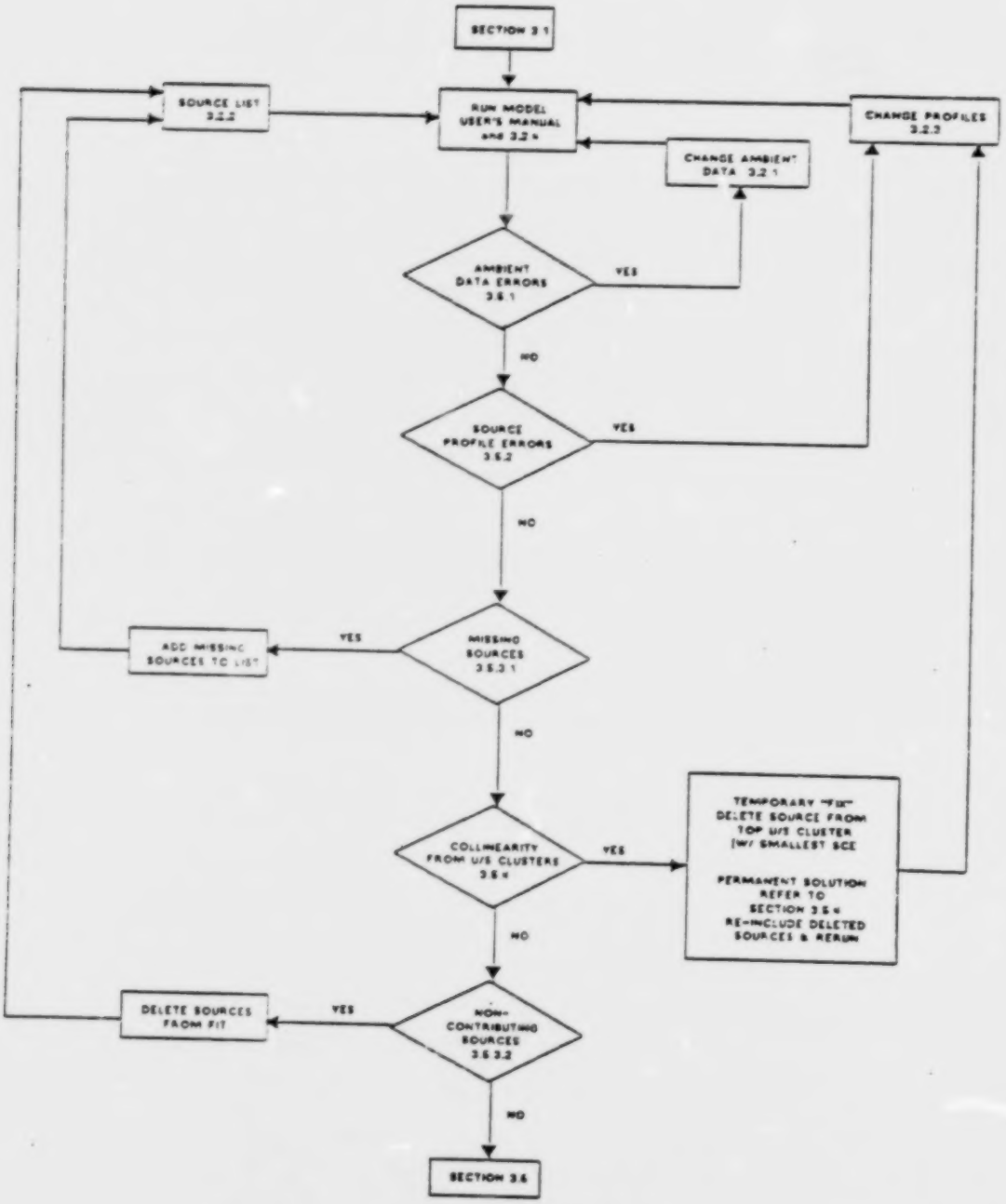


Figure B. Flowchart for Problem Identification And Correction

There are four main categories of situations that can be addressed to improve model performance. These are: (1) incorrect ambient data, (2) incorrect source profiles, (3) incorrect source list, and (4) profile uncertainty/collinearity. Indication of these situations is inferred from the diagnostics and indicators listed above. The indicators and appropriate corrective action for each situation is suggested. After corrective action has been taken, the model's fit of the measured species data is reevaluated. If the fit is acceptable for that particular use, the stability of the model results should be assessed. This includes evaluating the sensitivity of the model's results to errors in the sources, source profiles and the ambient data. Finally, the model results can be declared valid for that application.

If (a) the CMB model is determined to be applicable (Section 3.1), (b) the summary statistics and diagnostics are generally within target ranges (Section 3.3), (c) there are no significant deviations from model assumptions (Section 3.5), and (d) the sensitivity tests in Section 3.6 uncovered no unacceptable instability or consistency problems, the CMB analysis is considered valid.

TABLE A

SUMMARY OF MODEL DISPLAYS (OUTPUTS, STATISTICS, DIAGNOSTICS AND STATUS)

Output/Statistic/Code	Abbreviation	Description
Source Contribution Display		
Source Contribution Estimate	SCE	Primary output: the estimate, in ug/m^3 of each source or source category's contribution to the fitted ambient data.
Standard Error	STDERR	Primary output: indicator of the precision or uncertainty of the SCE, the variance of the SCE. [Target << SCE].
t-statistic	T-STAT	Statistic used to interpret the STDERR. It is the ratio of the SCE to its STDERR. A high T-STAT suggests a non-zero SCE. [Target > 2.0].
R-square	R-SQUARE	Statistic: used to measure the variance in the ambient species concentrations which is explained by the calculated species concentrations. Ranges from 0 to 1.0. [Target 0.8 to 1.0].
Chi-square	CHI-SQUARE	Statistic: similar to R-SQUARE except that it also considers the uncertainties of the calculated species concentrations. A high CHI-SQUARE suggests that the model has not explained the species data well. [Target 0.0-4.0].
Percent Mass Accounted For	PERCENT MASS or % MASS	Statistic: used to track the % of ambient mass (in the size fraction being analyzed) explained by the sum of the SCE's. A % MASS near 100% can be misleading because a poor fit can force a high % MASS. [Target 100% $\pm$ 20%].
Degrees of Freedom	DF	Statistic: no. species in fit - no. sources in fit. [Target > 5]
Site/sample	SITE, etc.	Status: data being used in the fit.

TABLE A (Continued)

SUMMARY OF MODEL DISPLAYS (OUTPUTS, STATISTICS, DIAGNOSTICS AND STATUS)

Output/Statistic/Code	Abbreviation	Description
Uncertainty/Similarity Display		
Uncertainty/Similarity Clusters	U/S CLUSTERS	Diagnostic: summary of singular value decomposition analysis. It shows clusters of sources which the model cannot easily distinguish between and that are likely to be interfering with the model's ability to provide a good set of SCE's. [Target - No Clusters].
Sum of Combined Sources	SUM ±	Supplemental Output: supplement to U/S Clusters. It estimates the sum of SCE's of the sources in a cluster and the std. error of the sum. Not needed if source profiles of cluster sources can be improved. The std. error of the SUM follows the ±.
Species Concentration Display		
In the fit	I	Status: * in column indicates species in fit.
Missing Measurement for Species	M	Status: M in column indicates missing measurement.
Measured Species Concentration	MEAS	Status: Ambient species concentrations (measurements and uncertainties) included in the data file.
Calculated Species Concentration	CALC	Output: the model's calculated concentration (and uncertainty) for each species in the fit.
Ratio of Calculated to Measured	RATIO C/M	Diagnostic: the simple ratio of CALC/MEAS and the uncertainty of the ratio. Used to identify species that are over/under accounted for by the model. [Target 0.5 to 2.0].

TABLE A (Continued)

SUMMARY OF MODEL DISPLAYS (OUTPUTS, STATISTICS, DIAGNOSTICS AND STATUS)

Output/Statistic/Code	Abbreviation	Description
Ratio of Residual to Its Uncertainty	RATIO R/U	Diagnostic: the signed difference CALC-MEAS divided by the uncertainty of that difference. Used to identify species that are over- or under-accounted for by the model. [Target |<2.0|].
Command Display		
Species - source contribution	SSCONT	Diagnostic: shows the fraction of each measured species concentration that is accounted for by the calculated species for each source or source category. This can be > 1.0 for a particular source if that species is over-accounted for by the fit. It is used to identify the sources which are accounting for particular species.

PROTOCOL FOR APPLYING AND VALIDATING THE CMB MODEL

1.0 INTRODUCTION

The Chemical Mass Balance receptor model (CMB) and Dispersion Models (DM) are two primary methods used to provide the source apportionment which forms the basis of control strategy development for PM_{10}. Other methods are allowed (US EPA 1987C). However, this document deals only with the application and validation of the CMB model.

The CMB model recommended by the EPA is a refinement of a model originally developed by Watson (1979) at the Oregon Graduate Center. This model has undergone five major revisions, with major contributions by Patrick Hanrahan and John Core of Oregon's Department of Environmental Quality; Hugh J. Williamson and Dennis A. DuBose of Radian Corporation; Luke Wijnberg of PEI Associates; John Watson, Norman Robinson, and Judith Chow of the Desert Research Institute; and Thompson G. Pace of the U.S. EPA. Version 6 of the CMB model, on which this protocol is based, was prepared and tested under a cooperative agreement between EPA's Office of Air Quality Planning and Standards and the Desert Research Institute of the University of Nevada. The software and protocols have been extensively tested by research, regulatory, and consultant users who have been identified in the acknowledgments. The Version 6 software is described in a User's Manual (U.S. EPA, 1987B).

The fundamental CMB model equations have been subjected to verification and evaluation using both real and simulated data as a part of the Quail Roost II Conference (Stevens and Pace, 1984). Additional verification and evaluation efforts have been undertaken by several investigators (Watson et al., 1984, DeCesar and Cooper, 1982, Dzubay et al., 1984, Gerlach et al., 1982, Currie et al., 1984; Watson and Robinson, 1984; Javitz and Watson, 1986; Watson and Chow, 1986; Henry and Kim, 1986). These efforts provide a basis for CMB applications in regulatory assessments of PM_{10} levels.

The CMB consists of a set of linear equations which express the ambient concentrations of chemical species as the sum of products of source compositions and source contributions. These equations are a direct consequence of the law of conservation of mass. This set of equations is overdetermined (more than one possible solution) because the number of measured chemical species generally exceeds the number of contributing source types. The source contributions are generally the unknowns in these equations. However, a unique solution cannot be found for this set of equations because measurement uncertainty precludes determination of exact values for source and receptor data. When these uncertainties are estimated for both source and receptor measurements, additional physical constraints can be applied which will yield a unique solution. This solution minimizes the difference between calculated and measured receptor concentrations by using a weighting scheme, effective variance. The weighting has a physical significance in that it is derived

from the measurement uncertainties of both source and receptor chemical species. Though the CMB solution is identical to some statistical inference methods, it is not dependent on statistical principles. The basic model equations which represent the source receptor relationship, the effective variance weighting, and the error propogation are all based on physical principles.

The CMB provides a source contribution estimate (SCE) and associated uncertainty (STDERR) for each source category that is chosen beforehand for consideration with the model. The model produces this estimate by making an effective variance weighted least squares "fit" between the chemical composition of the ambient sample and the composition of the sources. It estimates what amounts of each source (the SCE's) will collectively best "explain" the chemical composition of the ambient sample.

The objectives of this applications and validation protocol are to provide technical guidance on:

- The selection of source categories to be apportioned by the model and the selection of model input data.
- Methods to determine the validity and uncertainty of a specific model application.
- Methods to improve the quality and reduce the uncertainty of a specific model application.

This guidance consists of a seven-step procedure. This procedure is necessary because there is no single parameter or indicator which can assess the model's validity. The seven steps are presented below and discussed further in Section 3.0.

Section 1.0 serves as an introduction to this protocol and summarizes the steps which constitute the protocol. Section 2.0 discusses useful preliminary analyses. Section 3.0 presents detailed guidance associated with each of the seven steps specified above. Methods for assessing applicability of CMB to the problems at hand are discussed in Section 3.1. Initial choice of source categories and fitting species for use in CMB are addressed in Section 3.2. Section 3.3 describes the diagnostics and statistics which measure the performance of the model and identify possible deviations from model assumptions. Section 3.4 discusses the problems with model assumptions which are often encountered in an individual application.

Section 3.5 describes the rationale for identifying problems and making changes to the model inputs if statistics and diagnostics indicate that the model assumptions are not being adhered to. These problems include: incorrect ambient data; incorrect source profiles; incorrect source identification, uncertainty in source profiles and collinearity (similarity of

source profiles). Section 3.6 describes methods to test the consistency and stability of the model. Section 3.7 discusses the need to examine the model results with respect to the preliminary analyses and the model statistics to determine its validity. Appendix B includes examples and a "quick guide" for changing the model inputs.

Discussions of technical issues concerning the use of the DM and CMB models are found in the Guideline for Air Quality Models (Revised) (U.S. EPA, 1986A) and the Chemical Mass Balance User's Manual (U.S. EPA, 1987B). Guidance on the use of models in SIPs is discussed in U.S. EPA 1987C and guidance for reconciling differences in DM and CMB results is found in U.S. EPA 1987D.

2.0 PLANNING THE CMB ANALYSIS

2.1 Other Useful Receptor Methods

A CMB application is often preceded by a preliminary examination of existing data. This preliminary analysis is intended to identify the source types and chemical species to be included in the model. This in turn is used to support the physical basis for the sources included in the fit. It also provides the basis for determining the applicability of the CMB model to the situation under study and to the available data. Table 2.1 lists the receptor analysis methods which are the most useful. Further information on these methods can be found in U.S. EPA 1987C, and technical information can be found in the appropriate references listed in Table 2.1.

Another necessary activity which precedes the CMB is an estimate of PM_{10} background concentrations using the procedures in Appendix D of the PM_{10} SIP Development Guideline (U.S. EPA, 1987C).

2.2 Levels of CMB Analysis

After the CMB has been determined to be appropriate for a given situation, it can be applied with three sequential levels of complexity (U.S. EPA 1984A), each level being more costly, but supplying more accurate and precise information than the previous level. The levels are useful as a shorthand notation of the general level of comprehensiveness of a CMB study but have no regulatory significance. A given level may not provide valid results because of data limitations. In such cases, the next higher level may need to be undertaken to complete the CMB.

The basic level of CMB application (Level I) uses existing data or data that can be readily obtained from analyses of existing samples (Gordon, et al. 1984). This effort confirms the selection of contributing sources from the preliminary analysis and eliminates minor contributors from further scrutiny. If the sources contributing to the high concentrations of PM_{10} are apparent and sufficiently certain, no further work will be needed. Otherwise, this effort serves to reduce the areas to be studied in greater detail under an intermediate (Level II) analysis.

TABLE 2.1

USEFUL RECEPTOR METHODS

Analyses	Most Useful W/Suspected Sources	References
Chem Mass Bal (CMB)	Source categories	U.S. EPA 1987B
Factor Analysis (FA)	Sources unknown	U.S. EPA 1985B
Optical Microscopy (OM)	Biologicals, aged vs. new dust, fugitive process emissions	U.S. EPA 1983A
Scanning Electron Microscopy (SEM)	Inorganic (non-carbon) source(s)	U.S. EPA 1983A
X-Ray Diffraction (XRD)	Crystalline sources (e.g., fugitive dust)	U.S. EPA 1983A
Trajectory Anal. (TA)	Regional sources	Chow 1985
Emission Inventory	Any local source	U.S. EPA 1986A, 1987C
Carbon Analysis	Combustion or biological source	U.S. EPA 1985A
Sulfate Analysis	Regional Sources	U.S. EPA 1981C,D
Temporal Analyses	Seasonal or intermittent source	U.S. EPA 1978, 1984A
Spatial Analyses	Single source, line source	U.S. EPA 1978, 1984A, 1986A
Activity Patterns	Seasonal or intermittent source	U.S. EPA 1984A
Episode Days	Regional Sources, Exceptional Events	U.S. EPA 1984A, 1986
Source Profiles	Any local source	U.S. EPA 1985C
Background Analysis	Local vs. Regional Components	U.S. EPA 1987C
Meteorological Data	Supports DM, Spatial Analyses, FA	U.S. EPA 1984A, 1986A

Tracers (Simple CMB)	Source	Tracer	Size
	Auto	Pb	F
	Fuel oil	V	F
	Aluminum	F	F or C
	Burning Biomass	K	F or C
	Paint Pigments	Ti	C
	Galvanizing, refuse incin.	Zn	F
	Marine Aerosol	Na,Cl	C
	Fertilizer	P	C
	Resid. Oil	Ni, V	F

The intermediate (or Level II) analysis involves additional chemical analyses on existing samples or the acquisition of additional samples from existing sampling sites. It is intended to fill the gaps in model input data which may have been discovered in Level I so as to reduce uncertainty in results of the Level I source apportionment. A comprehensive CMB analysis (Level III) involves the acquisition of new data from new source and ambient sampling activities.

The CMB applications and validation protocol described here is appropriate to all three levels of PM_{10} assessment. It provides estimates of precision and validity which serve to define the measurement requirements for the next level of analysis. These estimates can also be used to determine whether or not the model results at a given level of PM_{10} assessment are certain enough to eliminate the need for more extensive assessment.

3.0 APPLICATIONS AND VALIDATION PROTOCOL

It is important that each application of the CMB be validated. The steps in a CMB analysis which will increase the assurance that the results are valid are:

(1) determine the general applicability of the CMB model to the application at hand;

(2) set up the model by identifying and assembling the source types, source profiles, and receptor concentrations needed for model input. Make a preliminary application of the model to these data;

(3) examine the model's statistics and diagnostics to identify potential deviations from the model assumptions;

(4) evaluate problems which might result from problems with model input data deviations from model assumptions;

(5) make any model input changes which can be justified to resolve the identified problems--rerun the model;

(6) assess the stability of the model results and their consistency with the preliminary analyses; and

(7) evaluate the model results by comparing them with other receptor or dispersion model results and reconcile any differences (U.S. EPA 1987D).

Successful execution of the CMB model using these seven steps will increase assurance that the CMB model is valid for the particular application and for individual cases within that application.

3.1 Determining Applicability of the CMB

The applicability of the CMB model must be determined for use with each specific ambient and source configuration. At this point, applicability is qualitatively determined. After the model is run, quantitative measures are available, as discussed in Section 3.4. The following conditions must be met for the CMB to be applicable:

(1) a sufficient number of PM_{10} receptor samples have been taken with an accepted sampling method to evaluate compliance with the annual and 24-hour standard (U.S. EPA 1987C) or to fulfill other objectives of the study.

(2) samples are amenable to or have been analyzed for a variety of chemical species. Minimal analyses include concentrations of Al, Si, S, Cl, K, Ca, Ti, V, Mn, Fe, Ni, Cu, Zn, Br, and Pb. Preferable additional analyses would include other elements such as Cr, As, and Se; elements; cations; anions; and elemental and organic carbon. Analytical methods should be chosen such that the concentrations of the majority of these species will exceed the detection limit and the variability of the filter blank. The uncertainties of these concentrations must also be available or estimable. Sampling and analysis procedures are discussed in greater detail by Gordon et al. (1984) and U.S. EPA (1981B);

(3) the potential source contributors can be identified and grouped into source categories of distinct chemical compositions with respect to the receptor species available from requirement 2. The degree of difference in source compositions necessary for successful CMB model application is discussed in Section 3.3.6;

(4) compositions for the source categories are obtainable which represent the source profile as it is perceived at the receptor for the chemical species available from requirement 2. Changes in source composition between source and receptor must be accommodated in order for the model to be physically meaningful. Appendix C of this protocol discusses the possible changes to the aerosol during transport and discusses methods to deal with these changes. Initial source profiles for Level I assessment may be obtained from the EPA source composition library (U.S. EPA 1985C). Whenever possible, these source profiles should be determined by actual sampling and analysis of sources in the airshed by appropriate methods (Chow et al., 1986, provide a summary of such methods); and

(5) the number of source types in a single application of the CMB must be fewer than the number of chemical species measured above lower quantifiable limits at the receptor.

Unless all five of the above requirments are met, the Chemical Mass Balance receptor model is not applicable to the situation under study. These are necessary, but not sufficient, requirements, and it may still be found that even though these requirements are met, the precision and validity

of CMB results are not adequate for control strategy decisions. The remaining steps in the applications and validation protocol must be taken to arrive at this conclusion, however.

3.2. CMB Model Setup and Initial Run

The CMB model is configured by selecting up to 35 chemical species and up to 35 source profiles for inclusion in the input data files. These are the prepared files from which a user selects species and sources to be included in a particular fit. The number of individual samples which may be examined in a single application is limited only by computer file storage constraints. Very few ambient data sets will contain measurements of more than 35 species in quantities above detection limits. A subset of these species and source profiles is used in each weighted least squares "fit" of the source profiles to the receptor data. A maximum of 21 species and 16 source profiles is allowed in an individual "fit" with the Version 6 software. The selection of which species and source profiles to include in the "fit" may vary from sample to sample. Both the source profile and receptor measurements must be of specified quality. Several other EPA documents offer guidance on data quality:

- The spatial and temporal representativeness of the data are consistent with U.S. EPA, 1987C and 1986B.
- The mass and chemical data used in the CMB should be subjected to thorough quality assurance procedures in accordance with an approved Quality Assurance Project Plan, prepared prior to the data collection (U.S. EPA, 1980).
- Measurement precision should be determined from replicate measurements as prescribed in the QA plan (e.g., Watson et al., 1983).

Quality assurance must be applied to source characterization measurements and to meteorological measurements as well as to receptor data. The CMB model can return results which are no more precise or valid than the data with which it is supplied. Incorrect data and uncertainty estimates are major causes of inconsistent CMB results. The applications and validation protocol identifies some (but not all) of the cases in which the input data are inadequate.

3.2.1 Ambient Species and Their Uncertainty

Each measured species provides information to help the model "fit" the data and arrive at the best of source contribution estimates. All available validated ambient species data (up to the 21 maximum allowed by the version 6 software) should be included as fitting species in the CMB. There may be reasons to change the ambient data or delete a species from the fit after the initial model run, as discussed in Section 3.5.

It is important that:

(1) values for all species are present in the source profiles and in the receptor data used in the fit. A default value of zero with a standard deviation equal to an analytical detection limit may be assigned to a species in a source profile if that species is known to be absent from that source type from previous tests of similar sources. If values are not assigned to a species in the source profile file, the Version 6 software automatically assigns a value of zero. This may bias the source contribution estimates if that species is actually present in a nonzero amount.

(2) only one of different measurements of the same species (such as elemental and total carbon or sulfur and sulfate) is included in the fit. If more than one measurement of the same species is included in the CMB solution, then that species influences the source contribution estimates more than it should; and

(3) species with values below detection limits may be included only if their uncertainty is also included. Minimum detection limits may be used to estimate this uncertainty if it is not otherwise reported. If the uncertainty is underestimated or is not specified (and given a default value of zero), then these very imprecise measurements will have an excessive influence on the source contribution estimates; and

(4) if a compound which is secondarily formed or is normally associated with regional scale pollution (such as sulfate) is included as a fitting species, a "single constituent source type" (after Watson, 1979) must also be included in the fit, unless that compound is felt to originate directly from sources which are included as fitting sources.

Uncertainties assigned to the measurements for use in the CMB should be reviewed to ensure that they are realistic estimates. Measurement uncertainties should be provided as part of the measurement process. Typical measurement uncertainties are on the order of +-5% to 20%, with some species being more uncertain than others because of analytical interferences and proximity to detection limits. Uncertainties in source profiles could be much greater, as discussed in Section 3.2.3. The model considers these uncertainties when it develops the "fit". Species with high uncertainties are unlikely to be very influential in the fit.

Chemical measurements on PM_{10} samples are usually reported with their measurement uncertainties which are normally determined via error propagation of chemical analysis and flow rate uncertainties (e.g., Watson et al., 1983). These uncertainties are determined from periodic performance tests and replicate analyses. The reporting of these uncertainties should be specified when the measurements are made. If chemical concentrations are

available without uncertainties, typical uncertainties may be assigned based on those reported in previous analyses (e.g., Mueller et al., 1983). The value of the diagnostics provided by the CMB software is substantially decreased without an adequate and accurate definition of measurement uncertainties in receptor data.

The individual samples should be run separately in the CMB, in most cases. Compositing or combining the data from several samples will usually decrease the number of sources that the CMB can resolve. Likewise, separate analysis of the fine and coarse ambient samples collected by a size fractionating sampler is preferable to analysis of a "total" sample which combines the two size fractions. The sources contributing to these two size fractions are generally quite different.

3.2.2 Source Identification - The Fitting Sources

A careful selection of sources for the initial run of the CMB will minimize errors in the fitting source list. Collinearity may be introduced if extra or incorrect sources are introduced. Section 2.0 summarized the types of preliminary analyses which can be carried out to assemble a model input file of up to 35 source profiles.

The following is a reasonable approach to selecting a set of up to 35 source types and profiles which can be stored in the source file:

(1) include ubiquitous area sources, such as motor vehicle exhaust, residual oil combustion, and resuspended dust. These sources are almost universally present in all urban areas;

(2) include natural sources, such as sea salt, if the receptor is in an area likely to be affected by such sources;

(3) include point sources which have been identified from an emissions inventory; and

(4) include "single constituent source types" (after Watson, 1979) in cases where substantial amounts of secondary nitrate, sulfate, and organic carbon are expected. These are profiles which represent only a single compound such as sulfate. They are used to account for that portion of the compound which is not accounted for by sources in the airshed. It is mainly used for secondary aerosol and sometimes carbon compounds for which a source cannot be readily identified. The profile contains only that compound and its fractional composition in the profile is set to 1.0. A compound which is secondarily formed or is largely emitted by sources outside of the airshed should generally not be included as a fitting species unless it is also included as a single constituent source type.

Several source profiles for each source type may be included in the source profile input files, but only one profile from each type should be included in a fit.

The following procedures can be followed to select a subset of these 35 source types for inclusion in the initial CMB "fit."

(1) review wind direction data and eliminate sources that have virtually no chance of contributing a detectable concentration because they are downwind of the receptor;

(2) based on the emission rates, stack height, distance from the receptor, microscopy and other preliminary analyses, eliminate those sources or source categories that are minor contributing sources [minor might be considered those sources in the airshed that individually account for less than 1 to 2% of the local source contribution to a receptor and are also minor contributors (< 5 to 10%) to each measured species concentration].

(3) eliminate those source types which are not likely to be emitting during the period of time being studied (e.g. woodsmoke emissions during hot summer months); and

(4) select only one source profile per source type (develop a weighted average profile, if necessary).

The final selection of the most appropriate source types and the profiles to represent those types results from interactive applications of the CMB with an evaluation of the diagnostic measures (see Section 3.3). It is possible that more than one subset of source types and source profiles will fit the receptor data equally well. The interactive application of the model to different source subsets will identify these cases.

3.2.3 Source Profile Representativeness and Uncertainty

The quality of the chemical measurements (elements or species) used in the source profiles is just as important as the ambient data quality. Source profile values must not only be accurately and precisely measured, they must also represent the range of variability expected from a number of individual emitters in the same source type category.

Some sources have emissions that are chemically similar or consistent over time - that is, although the absolute magnitude of the emissions may vary, the relative composition of many of the measured species present in a source may be sufficiently stable. However, the chemistry of some species could be variable if the source changes its operating conditions, feedstock or fuel. This variability must be reflected in the uncertainties that are assigned to each species in the profile. (These concerns about source profile variability are analogous to those faced by the dispersion modeler when estimating emission rates or dispersion parameters).

Because the CMB model uses the information provided by all species included in the fit, mis-estimation of a single species, even so-called "tracer" species, may not appreciably affect the source contribution estimates. This is especially true if these species have been assigned uncertainties which reflect their variability. When these uncertainties are adequately estimated, other, less variable species provide a larger influence on the source contribution estimates. Appendix A contains additional discussion on the source profiles and their uncertainties.

3.2.4 Initial Run of the CMB Model

Once the ambient data and source profiles and uncertainties have been assigned, it is time to make the initial model run. The user's manual (U.S. EPA 1987B) provides instructions for running the model. All selected sources and all fitting species are initially included. Following the initial run, the source contribution estimates (SCE) and the statistics and diagnostics will help to determine the validity of the model results. This step is similar to comparing the results of a dispersion model to measured data in that it provides feedback on the model's performance. Three output displays are provided: 1) the source contribution display; 2) the Uncertainty/Similarity display; and 3) the species concentration display. Also, several commands are available to provide the user with additional diagnostics and detail.

3.3 Outputs, Statistics and Diagnostics - Definitions and Interpretation

The primary outputs or variable values estimated by the CMB are: (1) source contribution estimates (SCE), (2) the std. error of the SCE, and (3) the calculated concentration for each species. These outputs are discussed in detail in the user's manual to the CMB software (U.S. EPA 1987B).

The CMB also provides several performance statistics and diagnostics that indicate how well the model "fits" the specified data and, also, what problems may be affecting the model outputs. They can also be used to identify data which may be in error and deviations from the model assumptions. These include: (1) percent of total measured mass accounted for; (2) R-Square and Chi Square; (3) t-Statistic; (4) Source Uncertainty Clusters; (5) RATIOS and residuals of the species; (6) Source-species contributions.

The following is a working summary of the meaning or interpretation of each diagnostic and performance statistic. A technical description of how each is calculated is presented in an Appendix to the CMB Version 6 User's Manual (U.S. EPA 1987B), and Appendix B to this document illustrates their use. Table 3.1 provides a useful reference summary of the outputs, diagnostics and statistics. Target values are suggested for each diagnostic/statistic. These rules-of-thumb are generally based on a combination of experience and theory.

3.3.1 Percent Total Mass Accounted For

The sum of the source contribution estimates should account for essentially all of the measured mass within the uncertainty of that sum and of the measured mass. Total mass is never used as a fitting species; thus, it is very useful as a goodness-of-fit indicator. Percentages outside of the range of 100 ± 20% could be caused by incorrect source profiles, incorrect ambient mass or species data, too many or too few source types, or lack of measurements for high concentration species. A % mass outside of 100 ± 20% coupled with a low R-Square and a high Chi-Square (see Section 3.3.2) is cause for concern. However, a % near 100 could be misleading because a poor fit (or collinearity) can "force" the % mass to be near 100%. Sources containing species which have high residuals (Section 3.3.5) may be misspecified. Low values for percent mass could be caused by unmeasured species. Also, primary emissions from combustion sources often contain substantial quantities of elemental or organic carbon, and their gaseous emissions often result in secondary particles such as sulfate or nitrate which comprise significant fractions of total mass. Measurement of these species and inclusion of single constituent source types as fitting sources can increase the value of the percent mass accounted for. [Target 100% ± 20%].

3.3.2 Chi-Square and R-Square -- Goodness of Fit

Both Chi-Square and R-Square statistics are indicators of the goodness-of-fit of the model estimates.

The value of Chi-Square is inversely proportional to the squares of the uncertainties in the source profiles and in the receptor data. A high Chi-Square (generally greater than 4.0) also suggests that the model has not adequately explained the values for the fitting species. A high Chi-Square may result if the source or ambient uncertainties have been underestimated, even though the fit is totally correct and is the best obtainable, given the true uncertainty of the input data. [Target 0.0 to 4.0].

R-Square is a measure of the variance in the ambient species data that is explained by the species values used in the fit. This is done by comparing the original species concentrations to those calculated from the model estimates. A low R-Square (generally less than 0.8) usually suggests that the model has done a poor job of explaining the receptor concentrations of the fitting species. This could be due to errors outside of the stated uncertainty intervals for source profiles or ambient data, or the poor fit could be due to "missing" sources (i.e., sources that should have been included in the fit but were not). If the percent of total mass accounted for is low when R-Square is low (see Section 3.4.2), then "missing" sources are possible causes.

TABLE 3.1

SUMMARY OF MODEL DISPLAYS (OUTPUTS, STATISTICS, DIAGNOSTICS AND STATUS)

Output/Statistic/Code	Abbreviation	Description
Source Contribution Display		
Source Contribution Estimate	SCE	Primary output: the estimate, in ug/m^3 of each source or source category's contribution to the fitted ambient data.
Standard Error	STDERR	Primary output: indicator of the precision or uncertainty of the SCE, the variance of the SCE. [Target << SCE].
t-statistic	T-STAT	Statistic used to interpret the STDERR. It is the ratio of the SCE to its STDERR. A high T-STAT suggests a non-zero SCE. [Target > 2.0].
R-square	R-SQUARE	Statistic: used to measure the variance in the ambient species concentrations which is explained by the calculated species concentrations. Ranges from 0 to 1.0. [Target 0.8 to 1.0].
Chi-square	CHI-SQUARE	Statistic: similar to R-SQUARE except that it also considers the uncertainties of the calculated species concentrations. A high CHI-SQUARE suggests that the model has not explained the species data well. [Target 0.0-4.0].
Percent Mass Accounted For	PERCENT MASS or % MASS	Statistic: used to track the % of ambient mass (in the size fraction being analyzed) explained by the sum of the SCE's. A % MASS near 100% can be misleading because a poor fit can force a high % MASS. [Target 100% $\pm$ 20%].
Degrees of Freedom	DF	Statistic: no. species in fit - no. sources in fit. [Target > 5]
Site/sample	SITE, etc.	Status: data being used in the fit.

TABLE 3.1 (Continued)

SUMMARY OF MODEL DISPLAYS (OUTPUTS, STATISTICS, DIAGNOSTICS AND STATUS)

Output/Statistic/Code	Abbreviation	Description
Uncertainty/Similarity Display		
Uncertainty/Similarity Clusters	U/S CLUSTERS	Diagnostic: summary of singular value decomposition analysis. It shows clusters of sources which the model cannot easily distinguish between and that are likely to be interfering with the model's ability to provide a good set of SCE's. [Target - No Clusters].
Sum of Combined Sources	SUM ±	Supplemental Output: supplement to U/S Clusters. It estimates the sum of SCE's of the sources in a cluster and the std. error of the sum. Not needed if source profiles of cluster sources can be improved. The std. error of the SUM follows the ±.
Species Concentration Display		
In the fit	I	Status: * in column indicates species in fit.
Missing Measurement for Species	M	Status: M in column indicates missing measurement.
Measured Species Concentration	MEAS	Status: Ambient species concentrations (measurements and uncertainties) included in the data file.
Calculated Species Concentration	CALC	Output: the model's calculated concentration (and uncertainty) for each species in the fit.
Ratio of Calculated to Measured	RATIO C/M	Diagnostic: the simple ratio of CALC/MEAS and the uncertainty of the ratio. Used to identify species that are over/under accounted for by the model. [Target 0.5 to 2.0].

TABLE 3.1 (Continued)

SUMMARY OF MODEL DISPLAYS (OUTPUTS, STATISTICS, DIAGNOSTICS AND STATUS)

Output/Statistic/Code	Abbreviation	Description
Ratio of Residual to Its Uncertainty	RATIO R/U	Diagnostic: the signed difference CALC-MEAS divided by the uncertainty of that difference. Used to identify species that are over- or under-accounted for by the model. [Target $\|< 2.0\|$].
Command Display		
Species - source contribution	SSCONT	Diagnostic: shows the fraction of each measured species concentration that is accounted for by the calculated species for each source or source category. This can be > 1.0 for a particular source if that species is over-accounted for by the fit. It is used to identify the sources which are accounting for particular species.

R-Square is mathematically related to the reduced Chi-Square as shown in Appendix A. It will generally approach a value of 1.0 when the reduced Chi-Square indicates a good fit. Generally speaking, the Chi-Square statistic is more revealing because it indicates how well the model has explained the receptor concentrations relative to the source and receptor measurement uncertainties. R-Square does not contain this relationship to measurement uncertainty. [Target 0.8 to 1.0].

3.3.3 t-Statistic

The standard error of the source contribution estimates is an important indicator of the precision, or certainty in the model estimates. The "t" statistic is used to interpret the standard error. Both statistics are found on the source contribution display. A "t" value of less than 2.0 is generally used to identify model estimates that are not significantly different from zero. Source estimates with small "t" values may indicate that the source is not contributing a quantity which exceeds the detection limts of the modeled system. A low t-Statistic also implies that the standard error estimates themselves are uncertain representations of the true standard error.

The presence of low t values for several sources within a run may also result from collinearities among the source profiles or imprecision in either the source profiles or the ambient data. Section 3.3.4 discusses a more definitive indicator of collinearity that is provided on the source uncertainty clusters display. The preliminary analyses are useful in corroborating whether a suspected source is a likely contributor at the monitoring site. Corroborative use of the results of preliminary analyses is described in Section 3.6. [Target $t > 2.0$].

3.3.4 Uncertainty/Similarity Clusters

The interaction of certain source profiles may lead to high standard errors (and low t-Statistics) for one or more of the sources. High standard errors associated with a particular source contribution estimate may be attributable to the profiles of other sources as well as to the uncertainty of its own profile. The uncertainty/similarity display helps to identify clusters (U/S clusters), consisting of other possible sources whose profiles may be interacting to cause high standard errors on one or more of the sources in the cluster. It also contains an estimate of the SUM of the SCE's (and the uncertainty of the SUM) for the sources in each cluster.

The uncertainty/similarity display identifies sources which may be collinear or whose source profile uncertainties are large, making it difficult for the model to distinguish between those particular sources. These difficult-to-distinguish sources are the "clusters" and the sources in the cluster are identified by number. Only those clusters which contain a source whose SCE is uncertain (i.e., a t-Statistic < 2.0) are displayed; other clusters may be present but they are not displayed because

the SCE's for their sources are acceptable as is. A Singular Value Decomposition (SVD) analysis is performed to determine these clusters and the results are summarized in this display. The clusters are listed in rank order such that the cluster which causes the largest overall uncertainty/similarity problem is listed first. The criteria for screening the SVD to identify clusters is discussed in the user's manual. The inherent uncertainty is used to make this ranking (after Henry, 1982).

Reducing the values of the standard errors of the SCE's is the ultimate goal once the correct sources have been included in the model run. The U/S clusters display identifies those sources whose profiles, if improved, would most substantially affect the model's ability to estimate SCE's with reduced standard errors. The sources in these clusters are candidates for improvement to the source profiles, either by more precise measurement of the profiles or by adding additional species to make the profiles less similar (less collinear). Further discussion of collinearity and source profile uncertainty can be found in Appendix A and Section C.3 of Appendix C. The calculations are documented in an Appendix to the CMB Version 6 User's Manual.

SCE's of sources within groups which are near the top of the list will have the highest standard errors or uncertainties. However, the estimate of the "SUM" of the SCE's of all of the sources in a group is known with more relative certainty than are the individual SCE's. The program computes this sum and a revised uncertainty for the sum and presents this information on the right side of the display. This sum may prove useful for a particlar application because of the reduction in uncertainty. The sum is equivalent to combining the source types in the uncertainty grouping into a single source type. Source type resolution is lost, but greater confidence in the sum of the source contributions is gained. [Target - no clusters].

3.3.5 Ratios and Residuals of Fitting Species

The RATIO C/M of the calculated (CALC) species mass to measured (MEAS) species mass is a convenient indicator of the magnitude of the residual. Ideally, it is equal to 1.0. A ratio >> 1.0 means that more mass for a given species was accounted for by the model than was measured on the ambient sample. This statistic is found on the species contribution display. [Target - 0.5 to 2.0].

The RATIO R/U statistic, also found on the species contribution display, is useful for interpreting the significance of the ratio of calculated to measured receptor species concentration. The residuals (R) are the signed difference between the CALC and the MEAS values for each species at the receptor location. The uncertainty (U) is the uncertainty in the estimate of the residual. If the absolute value of the residuals statistic exceeds 2.0, the residual is high enough to be of concern. [Target $|< 2.0|$].

A RATIO R/U >> 2.0 or << -2.0 for a species could be caused by the following:

1) incorrect ambient measurement of that species;*

2) incorrect source profiles for that species;*

3) inclusion of noncontributing source in the fit ("high positive" residual (> 2.0) only); or

4) absence of a contributing source from the fit ("nigh negative" residual (< -2.0) only).

*This would include underestimating the uncertainty of that species in either the ambient data or one of the source profiles.

Both the RATIOS C/M and R/U provide insight into the magnitude of the difference between the measured and calculated mass (the residual) for each species. The RATIO R/U is generally a more revealing indicator since it combines both the magnitude of the residual and the uncertainty in the estimate of the residual into a single measure.

3.3.6 Source-Species Contributions

The SSCONT command shows the percent contribution of each source to each species included in the fit. This information is used to identify potentially incorrect profiles or an incorrect source list which might be responsible for large species residuals (RATIO R/U). For example, if a species has a large residual and the SSCONT indicates that a particular source "accounts for" almost all of that species CALCULATED value, the profile value for that species should be carefully reviewed along with the ambient data for that species to determine which is in error.

3.4 Deviations from Model Assumptions

The CMB diagnostics and statistics discussed in Sections 3.3.1 through 3.3.6 indicate when deviations from model assumptions may have occurred. These deviations do not necessarily invalidate the CMB results-they merely indicate the potential for invalidity. This is why a separate step is necessary in the applications and validation protocol which evaluates the effects of these deviations from assumptions and determines whether or not these effects can be tolerated.

The CMB model is based on several assumptions and the model should be applied in a manner that is as consistent as possible with those assumptions. These assumptions are explicitly listed below and are further discussed in Appendix C to this document. Deviations from these assumptions may result in unacceptably large errors in the source contribution estimates. Therefore, these assumptions should be reviewed when the model is applied to ensure that expected deviations from them will not significantly bias the source contribution estimates. The assumptions of the CMB with an effective variance solution are:

1. Compositions of source emissions are constant over the period of ambient and source sampling.

2. Chemical species do not react with each other, i.e., they add linearly.

3. All sources with a potential for significantly contributing to the receptor have been identified and have had their emissions characterized.

4. The number of sources or source categories is less than the number of species.

5. The source compositions are linearly independent of each other.

6. Measurement uncertainties are random, uncorrelated, and normally distributed.

3.5 Identifying and Correcting Problems by Changing the Model Inputs

There are four main categories of problems which, once they have been identified, can be addressed to improve the performance of the model. The problem categories are: 1) incorrect ambient data; 2) incorrect source profiles; 3) incorrect source list; and 4) profile uncertainty/similarity. These are discussed briefly below and with a Quick Guide and examples in Appendix B. The following subsections discuss ways in which the model's diagnostics and statistics discussed in Section 3.3 may indicate possible problems with the model input, their possible causes and corrective action. Note that in some cases, not all "indications" must persist for a problem to be present. The more "indications" that persist, the more evidence of a problem. Because of the complex interactions of all of the data in a least squares estimate, the statistics or diagnostics may not always be adequate to conclusively isolate a problem with model input. Additional physical evidence is also very helpful. A flowchart is presented in Section 3.5.5 to provide a systematic approach to identifying and correcting problems.

3.5.1 Correcting the Ambient Data - Gross Errors

There may be inaccuracies in the ambient species that have not been uncovered in the routine data validation. If the data are "suspect" and there are no apparent data entry or analytical errors, the next step would be to eliminate the suspect species from the fit and rerun the model. Examine the changes in the estimates for each source. If the estimate changes by more than one standard error, and if the receptor concentration or a source profile value for the removed species is suspect, then either remeasure the species or use the SCE calculated without that species in the fit. Example B.1 in Appendix B illustrates the identification of incorrect ambient data.

INDICATION:

- o RATIO R/U << -2.0 for a species suggests either the ambient data are high or the profile data are low for that species;*
- o RATIO R/U >> 2.0 for a species would imply that the ambient data are low or the profile data for that species are high. If profile data are suspect, see Section 3.5.2 and problem 8.2.*

ACTION:

1. Review the uncertainty assigned to the species with the high residuals. Make any justifiable and appropriate changes and rerun the CMB. If this improves the RATIO R/U, Step 2 is not necessary.
2. Delete the suspect species from the list of fitting species and rerun. If the SCE changes by at least one standard error, do not use this species in the fit until it has been remeasured.

* (NOTE: RATIO R/U can also indicate an incorrect source list - see Section 3.5.3; also, it can be due to an underestimated uncertainty for that species in either the ambient data or one of the source profiles.

3.5.2 Correcting Source Profiles - Gross Errors

A gross error in the value of one or more species in a profile might result in a high standard error in the SCE and a high residual for those species. Therefore, one or more high residual values suggests that the uncertain source profile (and the associated species in particular) be checked and remeasured if necessary. The high residual is a likely species to check for errors. Appendix B contains Example B.2 which illustrates the problem of gross errors in a profile. Indications of the problem are given below.

INDICATION:

- o SCE that is inconsistent with preliminary analyses or physical evidence;
- o one or more species has a "high (pos. or neg.)" residual which cannot be attributed to incorrect ambient data; further evidence of species error if the SSCONT reveals that one source contribution dominates that species.

ACTION: Review profile data for the suspect species carefully. Correct or remeasure profile if necessary.

3.5.3 Changing the Source List

3.5.3.1 Missing Sources

Missing source types are identified by a low percent mass explained (e.g., less than 80%) and/or a RATIO R/U << -2.0 for chemical species which are in the missing source. A "high negative" residual for one or more species and a high Chi-Square are also indicative of missing sources. The key to identifying these sources resides in the calculated to measured chemical concentrations listed in the SPECIES CONCENTRATION display. "High negative" residuals imply that a source is needed which will supply a larger quantity of that species. The PMATRIX command lists all of the source profiles in the model's input data file. These profiles can be examined to determine which ones would would supply sufficient quantities of the missing concentrations if they were added to the set of fitting sources. The CMB can be reapplied as many times as is necessary to determine which source types and source profiles best account for the underestimated receptor concentrations. A source should not be included in the final fit just because it "explains" the data; however, there must be a physical justification for the source's contribution at a receptor if it is to be included in the fit.

The source list can be changed by adding or deleting sources. Example B.3 illustrates the identification of a missing source. Indication of a missing source is given by the following conditions.

INDICATION:

- o High Chi-Square;
- o Low percent mass explained;
- o RATIO R/U << -2.0 (a "high negative" residual) for one or more species that are known to be present in the suspect source.

ACTION: Add source profiles to the fit and reevaluate.

3.5.3.2 Noncontributing Sources

Noncontributing source types, or better stated, source types with contributions lower than detection limits, are identified by T-STAT values below 2. Such source types may be eliminated from the fit if the source contribution is indeed small.

Noncontributing source situation is illustrated in Appendix B in Example B.5 and Indications are summarized below.

INDICATION:

- o T-STAT between -2.0 and 2.0
- o RATIO R/U >> 2.0 ("high positive") residual for a species which is attributed to the suspect source by the SSCONT diagnostic

o Negative SCE

o Physical basis for the source's contribution is weak.

ACTION: Delete source from fit.* However, if the source's contribution (SCE) is large and there is a strong physical basis for its presence, the profile should be remeasured to reduce its uncertainty.

*NOTE: If the source is present but a very small contribution to total mass, it should only be removed from the fit if the SSCONT shows that none of the species in the source account for more than 5 to 10% of the ambient concentration for those species.

3.5.4 Improving Source Profiles---Uncertainty/Similarity

As discussed in Appendix C.3, there are two reasons (other than gross errors) to improve the source profiles: (1) high profile uncertainty; (2) collinearity with low profile uncertainty. This section discusses methods of identifying collinear sources and ways to reduce the uncertainty in SCEs.

The uncertainty/similarity display identifies those source types which interact sufficiently to contribute to large standard errors in the source contribution estimates in that group. An indication that the interaction may be significant can be seen by noting the sources on the U/S cluster display.

A simple test is proposed to determine if the uncertainty in the SCE is due to high profile uncertainty: reduce the uncertainties in the profile to levels that might be reasonable to achieve if the source profiles were measured more precisely; then, rerun the CMB - if the clusters containing those sources are no longer listed, it is likely that collinearity per se is not significant. Remeasurement of the profile will probably improve the uncertainties of the source contribution estimates. It is possible that reducing the uncertainty will not eliminate the clusters but the SCE uncertainty will likely be improved somewhat. This would suggest that collinearity is also present. Appropriate action is discussed below. Example B.4 illustrates this problem.

INDICATION:

o Two or more sources listed in a U/S cluster

o T-statistic < 2.0 for one or more sources in that cluster-- if the T-STAT becomes > 2.0 when species uncertainties for profile for that source is arbitrarily reduced to a potentially achievable level, this indicates that the uncertainty in the source profile is at least partially responsible for the "apparent" collinearity.

ACTION: Remedies for unacceptably high uncertainties due to collinearity can take five forms ranked from most to least desirable.

(1) the profile of one or more of the cluster sources could be improved by measuring additional species.

(2) Reduce the uncertainties in the source profiles of the cluster sources. If the T-STAT becomes > 2.0, and if these profile uncertainties are realistically achievable by remeasurement, then the "apparent" collinearity can be improved in large part by improving the uncertainty in the profiles. Ideally, the U/S cluster for that group of sources would disappear. Remeasure and rerun the CMB with the improved measurements. More precise source profile measurements must be obtained before re-applying the model.

(3) The estimate of the SUM of the source categories in the U/S cluster can be used. Obtain independent estimates of the contributions of the individual source categories and use them to apportion the SUM into the source categories.

(4) combine the profiles of the collinear source profiles into a single signature of a "composite source category" that chemically represents the source categories identified by the U/S cluster. For example, resuspended road dust and windblown soil dust are chemically similar, and some modelers include a single term to represent "crustal material" instead of the two individual source types. This would result in improved source estimates of the crustal component, which can then serve as an estimate of the combined impact of the two sources. This aggregated estimate might then be partitioned into its components by another method (e.g., dispersion modeling, microscopy, or wind trajectory analysis).

(5) Species which are causing the similarity in source profiles might be deleted from the fit. These species can often be determined from the display produced by the SSCONT command. Often one of the cluster sources will be >> 100% for that species and the other will be negative. Unfortunately, eliminating too many species from the fit may cause the model to fail the applicability requirements in Section 3.1. Also, the results should acknowledge that the deleted source may be present.

3.5.5 Problem Identification and Correction Strategy

Figure 1 shows the order in which the above actions should be taken. Generally, it is best to make only one change at a time to the model setup before rerunning. An exception is that data errors should all be corrected when they are identified. This stepwise procedure may necessitate cycling through the steps several times. Each time a change is made, it may "clarify" the need to make a change that was not evident on a previous iteration (e.g., you may address a collinearity problem and reveal a data problem previously unidentified because of the collinearity). The cycling process is repeated until no changes are justified by the criteria in Section 3.5.1 through 3.5.4. Example B.6 in Appendix B illustrates a solution where multiple problems are present after the initial run. The use of this flowchart will greatly increase the consistency of CMB application among users. However, some operator judgments are necessary regarding data validity and corrective action(s) to address collinearity.

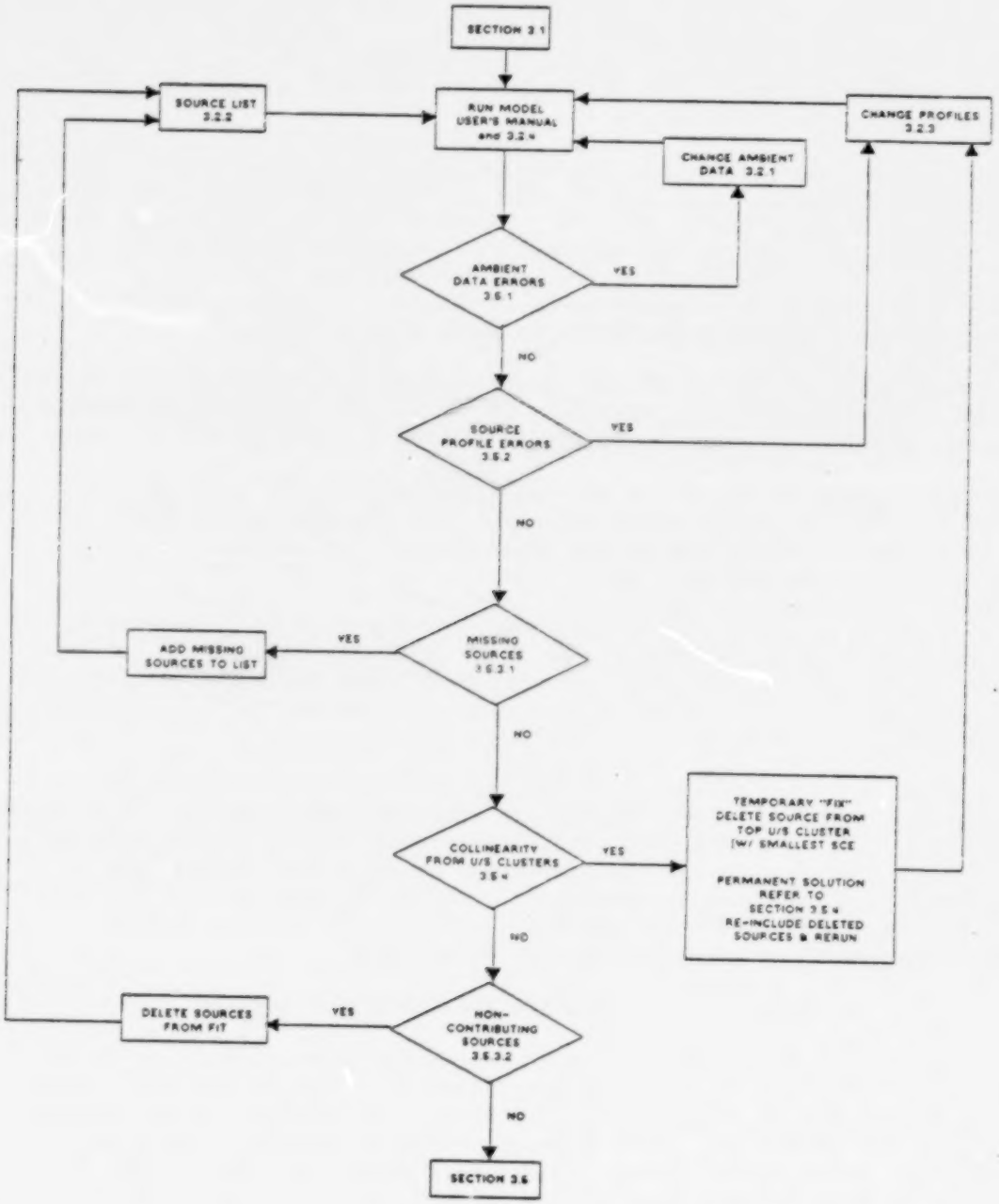


Figure 1. Flowchart for Problem Identification And Correction

3.6 Consistency/Stability of the Model Results

The CMB source contributions should be compared to the results of other receptor methods to provide corroboration of results. In addition to providing corroboration of CMB estimates, qualitative receptor analyses can provide clues to the causes of unresolved issues in the sensitivity and summary statistics reviews. For instance, meteorological data and spatial emissions inventories could be used to identify potential missing sources of a particular fitting element, as could microscopy or factor analysis.

In the event that significant inconsistencies are observed, all results should be reviewed, focusing on those particular sources that appear to be inconsistent among the different methods. If there is compelling evidence that the CMB model inputs be changed, the deviations from model assumptions, model sensitivity and summary statistics (Sections 3.3, 3.4 and 3.5) should be reevaluated in addition to rerunning the CMB model.

The CMB estimates should be tested to see how sensitive they are to the various input data. Unstable estimates (source contribution estimates that change by more than one standard error estimate) are an indication that the model may not be providing stable results. For CMB validation, model stability tests are usually taken to mean the evaluation of model estimates to changes in input parameters, such as the selected sources and their profiles, as well as selection of fitting species used to reach a solution with the CMB model. The following is a discussion of three types of parameter changes that should be included in a model stability (sensitivity) test, and a discussion of ways to make the model less sensitive (more stable).

3.6.1 Source Profile Sensitivity

The CMB model's effective variance fitting procedure uses estimates of the source profile and receptor concentration uncertainties to "weight" their effect in arriving at source contribution estimates. It is helpful to explore how sensitive the source contribution estimates are to changes in the source profiles and these uncertainties. This can be done by introducing changes into the source profiles and rerunning the model for each change.

The model user can select several species from a source(s) of particular regulatory interest and assign worst case values to those species in the profile. The model can then be rerun with the worst case profile(s). A practical way to accomplish this sensitivity analysis is to include a "worst case" source profile along with the "best estimate" profile in the "FS" or "CS" data file. The resulting source estimate(s) can be considered "brackets" to the source contribution estimates and can be compared to the uncertainty intervals calculated for each run. If the bracketing interval is greater than the calculated uncertainty interval, then the model may be sensitive to changes in the source profiles.

3.6.2 Receptor Concentration Sensitivity

The stability of source contribution estimates with respect to receptor concentrations is best tested with collocated chemical measurements from one of the sampling sites. These collocated measurements are usually included as part of the quality assurance plan for a subset of all samples. If nearly equivalent source contribution estimates are derived from these two independent measurements of the same ambient air, then the receptor data are not likely causing instabilities in the CMB results.

Lacking these collocated data, portions of the input data may be perturbed randomly or systematically in proportion to their uncertainty (e.g., Javitz and Watson, 1986; Watson and Robinson, 1984). The source contribution estimates for the sources of regulatory interest should not change by more than one standard error in response to small perturbations if the results are stable. (A "small" perturbation is defined as one std. error of the ambient species concentrations.) If the results are not stable, the validity of the CMB result for that particular data are questionable.

3.6.3 Fitting Species Sensitivity

The stability of CMB model results to the fitting species can be evaluated by identifying a species which SSCONT attributes in large part to a single source. Eliminate this species from the fit and examine how much the corresponding source contribution changes. If this change is greater than the STDERR, then that species must be greatly influencing the "fit." Review the quality of both the source and ambient measurements for that species carefully because of its influence on the model estimates.

3.7 Evaluating Results of the CMB Analyses

If (a) the CMB model is determined to be applicable (Section 3.1), (b) the summary statistics and diagnostics are generally within target ranges (Section 3.3), (c) there are no significant deviations from model assumptions (Section 3.5), and (d) the sensitivity tests in Section 3.6 uncovered no unacceptable instability or consistency problems, the CMB analysis is considered valid. If uncertainties associated source estimates are too high for decision-making purposes even after taking the steps recommended in this protocol, then the source compositions being used are not representative of the sources in the airshed, or they contain too much uncertainty associated with the influential species.

It is recommended that both a dispersion model and receptor model be used in a collaborative manner to perform an apportionment, provided that the dispersion model is applicable and the receptor model is valid for the particular application (U.S. EPA 1987C).

4.0 ACKNOWLEDGMENTS

The primary authors of this document were Thompson G. Pace, P.E., of the U.S. EPA and Dr. John G. Watson of the Desert Research Institute. However, the experience, talent and ideas of the co-authors played a large role in its development. The co-authors participated in a workshop to review and revise the document in May 1986 in San Francisco. The co-authors are: Dr. Kit Wagner, Mr. Bart Croes, Mr. Duane Ono, Dr. Hal Javitz, Mr. Mike Naylor, Dr. Judith C. Chow, Dr. David Maughan, Dr. Edwin Meyer, Mr. Luke Wijnberg, Mr. Ken Axetell, Mr. Mike Anderson, Mr. John Core, Mr. Pat Hanrahan, Dr. Ron Henry, Mr. Bong Mann Kim, Mr. Chung Liu, and Dr. Andy Gray.

Several others provided review and comment outside of the workshop. The reviews by Dr. Glen Gordon, Mr. William Cox, Mr. Chuck Lewis, Dr. Thomas D. Dzubay, Mr. Robert K. Stevens, Mr. Johnnie Pearson, Dr. Richard DeCesar, Dr. Sylvia Edgerton, and Ms. Bridget Landry are greatly appreciated. The typing and revisions by Ms. Cathy Coats, Ms. Jo Harris, and Ms. Linda Ferrell are much appreciated.

5.0 References

Blumenthal, D., Watson, J. G., Richards, L. W., Itering, S. V., and Chow, J. C. (1986), "Southern California Air Quality Study: Suggested Program Plan," Sonoma Technology, Santa Rosa, CA.

Chow, J. C., 1985, "A Composite Modeling Approach to Assess Air Pollution Source/Receptor Relationships," Doctor of Science Dissertation, Harvard University, Boston, MA, July 1985.

Chow, J. C., Watson, J. G., Egami, R. T., Wright. B., Ralph, C., Naylor, M., Smith, J., and Serdoz, R. (1986), "Program Plan for State of Nevada Air Pollution Study (SNAPS)," DRI Document 8086-100, Energy and Environmental Engineering Center, Desert Research Institute, NV 89506.

Chow, J. C., Watson, J. G., and Frazier, C. A. (1986A), "A Survey of Existing Fugitive/Area Source Characterization Methods for Receptor Modeling" to be printed in Proceedings, Particulate Matter Fugitive Dusts-Measurement and Control in Western Arid Regions, Air Pollution Control Association, Tucson, Arizona, Oct. 16-17, 1986.

Currie, L. A., Gerlach, R. W., Lewis, C. W. et al. (1984), "Interlaboratory Comparison of Source Apportionment Procedures: Results for Simulated Data Sets," Atm. Env., 18, (8):1517-1537.

Currie, L. A., Gerlach, R. W., Lewis, C. W., Balfour, W. D., Cooper, J. A., Dattner, S. L., DeCesar, R. T., Gordon, G. E., Heisler, S. L., Hopke, P. K., Shah, J. J., Thurston, G. D., and Williamson, H. J. (1984), "Interlaboratory Comparison of Source Apportionment Procedures: Results for Simulated Data Sets," Atm. Env., 18, 1517.

DeCesar, R. T., and Cooper, J. A. (1982), "Evaluation of Multivariate and Chemical Mass Balance Approaches to Aerosol Source Apportionment, Using Synthetic Data and an Expanded PACS Data Set," Proceedings, Receptor Models Applied to Contemporary Pollution Problems, Specialty Conference, Air Pollution Control Association, Denver, MA.

Dzubay, T. G., Stevens, R. K. et al. (1982A), "Intercomparison of Results of Several Receptor Models for Apportioning Houston Aerosol", Proceedings, Receptor Models Applied to Contemporary Pollution Problems, Specialty Conference, Air Pollution Control Association, Denver, CO, 1982.

Dzubay, T. G., Stevens, R. K. et al. (1984), "Interlaboratory Comparison of Receptor Model Results for Houston Aerosol," Atm. Env., 18 (8):1555-1566.

Friedlander, S. V. (1981), "New Developments in Receptor Modeling Theory," Atmospheric Aerosol: Source/Air Quality Relationships, ed. by E. S. Marias and P. K. Hopke, CS Symposium Series 167, American Chemical Society, Washington, DC, p. 1-p. 19.

Gerlach, R. W., Currie, L. A. and Lewis, C. D. (1982), "Review of the Quail Roost II Receptor Model Simulation Exercise," Proceedings, Receptor Models Applied to Contemporary Pollution Problems, Specialty Conference, APCA, Denver, MA.

Gordon, G. E. (1984), "Atmospheric Tracers of Opportunity from Important Classes of Air Pollution Sources," DOE Workshop on Atmospheric Tracers, Santa Fe, NM.

Gordon, G. E., Pierson, W. R., Daisey, J. M., Lioy, P. J., Cooper, J. A., Watson, J. G., and Cass, G. R. (1984), "Consideration for Design of Source Apportionment Studies," Atm. Env., 18:1567-1582.

Gordon, G. E. and Ann Sheffield, University of Maryland, Chemistry Building, College Park, Maryland, 20742, Personal communication to Tom Pace, April 1987.

Henry, R. C. (1982), "Stability Analysis of Receptor Models That Use Lowest Squares Fitting" in Receptor Models Applied to Contemporary Pollution Problems, edited by S. L. Dattner and P. K. Hopke, Air Pollution Control Association, Pittsburgh, PA, p. 141.

Henry, R. C. and Kim, B. M. (1986), "Evaluation of Receptor Model Performance," Report to U.S. Environmental Protection Agency, Technology Development Section, Air Management Technology Branch, Monitoring and Data Analysis Division, Office of Air Quality Planning and Standards, Research Triangle Park, NC.

Hopke, P. K. (1985), Receptor Modeling in Environmental Chemistry, John Wiley and Sons, New York, 1985.

Javitz, H. S. and Watson, J. G. (1986) "Feasibility Study of Receptor Modeling for Apportioning Utility Contributions to Air Constituents, Deposition Quality and Light Extinction," Draft Report for Electric Power Research Institute, prepared by SRI International, Menlo Park, CA.

Lewis, C. W. and Stevens, R. U. (1985) "Hybrid Receptor Model for Secondary Sulfate from an SO_2 Point Source," Atmospheric Environment, 19, 917-924.

Liu, C. S., Gray, H. A., Grisinger, J. E., and Davidson, A. (1986), "Draft 1987 AQMP Revision Working Paper No. 2: PM_{10} Modeling Approach," prepared by the South Coast Air Quality Management District, El Monte, CA.

Mueller, P. K. and Hidy, G. M. et al. (1983), "The Sulfate Regional Experiment: Report of Findings," Electric Power Research Institute Report #EA-1901, Palo Alto, CA.

Scheff, P. and Wadden, R. A. (1986), "Predicting Unidentified and Secondary Sources With Chemical Mass Balance Receptor Modeling," Receptor Methods for Source Apportionment--Real World Issues and Applications, Edited by T. G. Pace, APCA, Pittsburgh, PA, 1986.

Stafford, M. A. and Liljestrand, H. M. (1984), "On the Distinction of Secondary Species in Acid Deposition," presented at the 77th Annual Meeting of the Air Pollution Control Association, San Francisco, CA.

Stevens, R. K. and Thompson G. Pace (1984), "Review of the Mathematical and Empirical Receptor Models Workshop (Quail Roost II)." Atm. Env. 18, 1499-1506.

Trijonis, J., "Model Reconciliation," Special Report prepared for U.S. EPA and TRC Environmental Consultants under Contract 68-02-3886, Work Asignment No. 13, Research Triangle Park, NC 27711, September 1985.

U.S. EPA, 1978, Digest of Ambient Particulate Analysis and Assessment Methods, EPA-450/3-78-013, U.S. EPA, Research Triangle Park, NC 27711, September 1978.

U.S. EPA, 1980, Interim Guidelines and Specifications for Preparing Quality Assurance Project Plans, QAMS-005/80, U.S. EPA, Office of Research and Development, Research Triangle Park, NC 27711, December 1980.

U.S. EPA, 1981A, Receptor Model Technical Series, Volume I: Introduction to Receptor Models, EPA-450/4-81-016a, U.S. EPA, Research Triangle Park, NC 27711, July 1981.

U.S. EPA, 1981B, Receptor Model Technical Series, Volume II: Chemical Mass Balance, EPA-450/4-81-016b, U.S. EPA, Research Triangle Park, NC 27711, July 1981.

U.S. EPA, 1981C, "SOP for Technicon Determination of Sulfate in Suspended Particulate Matter Collected on Glass Fiber Filters," EMSL/RTP SOP-EMD-005, U.S. EPA, Research Triangle Park, NC 27711, November 1987.

U.S. EPA, 1981D, "SOP for the Extraction of Sulfate and Nitrate and for the Technicon Determination of Sulfate or Suspended Particulate Matter Collected on Dichotomous Filters," EMSL/RTP SOP-EMD-006, U.S. EPA, Research Triangle Park, NC 27711, November 1981.

U.S. EPA, 1983A, Receptor Model Technical Series, Volume IV: Summary of Particle Identification Techniques, EPA-450/4-83-018, U.S. EPA, Research Triangle Park, NC 27711, June 1983.

U.S. EPA, 1984A, Receptor Model Technical Series, Volume V: Source Apportionment Techniques and Considerations in Combining Their Use, EPA-450/4-84-020, U.S. EPA, Research Triangle Park, NC 27711, July 1984.

U.S. EPA, 1984B, Interim Procedures for Evaluating Air Quality Models, EPA-450/4-85-023, U.S. EPA, Research Triangle Park, NC 27711, September 1984.

U.S. EPA, 1985A, Technical Support Document for Residential Wood Combustion EPA-450/4-85-XXX, U.S. EPA, Research Triangle Park, NC 27711, June 1985, Draft.

U.S. EPA, 1985B, Receptor Model Technical Series, Volume VI: Multivariate Methods, EPA-450/4-85-007, U.S. EPA, Research Triangle Park, NC 27711, July 1985.

U.S. EPA, 1985C, Receptor Model Source Composition Library, EPA 450/4-85-002, U.S. EPA, Research Triangle Park, NC 27711, November 1984.

U.S. EPA, 1986A, Guideline on Air Quality Models (Revised) U.S. EPA, Research Triangle Park, NC 27711, revision in preparation.

U.S. EPA, 1986B, Procedures for Estimating Probability of Nonattainment of a PM_{10} NAAQS Using Total Suspended Particulate or PM_{10} Data, EPA-450/4-86-017, U.S. EPA, Research Triangle Park, NC 27711, December 1986.

U.S. EPA, 1986C, Guideline on the Identification and Use of Air Quality Data Affected by Exceptional Events, EPA-450/4-86-007, U.S. EPA, Research Triangle Park, NC 27711, July 1986.

U.S. EPA, 1987B, Receptor Model Technical Series, Vol. III (Revised), Chemical Mass Balance Receptor Model User's Manual, EPA-450/x-xx-xxx, U.S. EPA, Research Triangle Park, NC 27711, May 1987.

U.S. EPA, 1987C, PM_{10} SIP Development Guideline, U.S. EPA, EPA 450/2-87-001, Research Triangle Park, NC 27711, January 1987.

U.S. EPA, 1987D, Procedures for Reconciling Differences in Receptor and Dispersion Models, U.S. EPA, EPA 450/4-87-008, Research Triangle Park, NC 27711, May 1987.

Watson, J. G. (1979), "Chemical Element Balance Receptor Model Methodology for Assessing the Sources of Fine and Total Suspended Particulate Matter in Portland, Oregon," Ph.D. dissertation, Oregon Graduate Center, Beaverton, Oregon.

Watson, J. G., Lioy, P. J. and Mueller, P. U. (1983), "The Measurement Process: Precision, Accuracy and Validity," in Air Sampling Instruments for Evaluation of Atmospheric Contaminants, Sixth Edition, American Conference of Governmental Industrial Hygienists, Cincinnati, OH.

Watson, J. G. and Robinson, N. F. (1984), "A Method to Determine Accuracy and Precision Required of Receptor Model Measurements," APCA/ASQC Specialty Conference on: Quality Assurance in Air Pollution Measurements. Boulder, CO.

Watson, J. G. and Robinson, J. F. (1984), "A Method to Specify Measurements for Receptor Models," Proceedings of the National Symposium on Recent Advances in Pollutant Monitoring of Ambient Air and Stationary Sources, U.S. EPA, Research Triangle Park, NC, April, 1984.

Watson, J. G., Cooper, J. A., and Huntzicker, J. J. (1984), "The Effective Variance Weighting for Least Squares Calculations Applied to the Mass Balance Receptor Model," Atm. Env., 18:1347-1335.

Watson, J. G. and Chow, J. C. (1986), "Volume 4 - An Evaluation of Ambient Aerosol Chemistry in the Western United States," Draft Report of Western States and Deposition Project, Phase I of SYSAPP-86-129, prepared for Western Governor's Association, Denver, Colorado.

APPENDIX A

Determining Source Profiles and Their Uncertainties

Source profile variability needs to be assessed on a case-by-case basis by reviewing the nature of the fuels and raw materials used by the sources. Also, values and uncertainties of those species which are the most influential in the "fit" should be carefully examined to ensure that both the ambient data and source profiles are accurate, valid and precise. The SSCONT command (see Section 3.3.6) shows the percent contribution of each source to each species in the fit. This provides some indication of each species' influence on the fit.

Most combustion processes use fuels with variable chemical compositions. Coal and fuel oil combustion source profiles may change substantially when the origin of the fuel changes (e.g., South American vs. Middle East Crude). Process upsets and control device malfunctions also affect source profiles, but usually not as much as they affect the mass emission rates. Manufacturing process emissions may be stable over time, but the use of different raw materials or changing product specifications could cause source profiles to vary.

The most accurate and reasonably "certain" information about sources in an airshed is derived from samples of the effluents from those sources, preferably at the same time that the receptor measurements are being taken. This may only be possible during a comprehensive (Level III) CMB application. The alternatives are to measure the source profiles from representative sources at representative times in the airshed under study, or to use profiles measured on similar sources in other airsheds and compiled in a source composition library (U.S. EPA, 1985C).

Source sampling to obtain source profile information is not as complicated as sampling to determine the mass emission rate. This is because source profiles require only that species be characterized as % of total mass of sample collected. (Not in terms of their absolute emission rate.) Hot exhaust sampling, diluted (cooled) exhaust sampling, airborne plume sampling, ground-based plume sampling, and grab sampling methods have been developed and are applicable to different source types (Chow et al., 1986A).

"Grab sampling" is an especially useful and inexpensive method of characterizing soil dust or storage piles. The procedure is so named because samples of the effluent or raw material are "grabbed" in bulk instead of sampled from an exhaust stream. These sources can be sampled in bulk and resuspended onto filters in the laboratory so that chemical analysis can be performed. In some cases, it can be determined that the effluent captured by the source's control device is chemically representative of the emissions. In these cases, a bulk sample of the captured effluent may be likewise resuspended. The "grab sample" procedure is generally much less expensive than stack sampling and can yield acceptable profiles for some types of sources.

Any source profile contains both measurement and analytical uncertainty. Use of any profile data which are not representative of the source at the time the ambient sample was collected introduces additional (location and/or time) uncertainty in the source profiles. The combination of these (the overall uncertainty) is assigned to each species in a source profile and the CMB model uses these uncertainties to weight a species' influence on the solution. The uncertainty estimates for each species in the source profile are just as important as the source profile values themselves. Uncertainty estimates can be based on the judgment of an engineer or scientist knowledgeable about each particular source and its operating characteristics. This judgment takes advantage of natural lower and upper limits to the possible uncertainty. The analytical uncertainty derived from error propagation provides a lower limit of the uncertainty estimate in a source profile. An upper limit is imposed by the constraint that the sum of all the fractional chemical compositions in a source profile cannot exceed unity, and random uncertainties cannot be so large that this might occur. A value between these extremes is often appropriate for CMB analysis. Repeated source tests over a range of operating variables provides a better estimate of the average profile of a source type and the standard deviation of that average is a good estimate of its uncertainty.

The EPA Source Composition Library presents typical compositions and uncertainties for the most common source types. These uncertainties may be understated when the profile is used to represent conditions at another location. Uncertainties of magnitude similar or even greater than those in the Library should be assigned when source composition data without uncertainties has been obtained. The value of the CMB performance diagnostics is substantially reduced without accurate estimates of source profile uncertainty.

Previous applications of the CMB model suggest that typical uncertainties in both receptor and source measurements of up to +-30% in each of the species is tolerable; large uncertainties of +-50 to 100% or more for some species may also be tolerable (Watson 1979). As noted, the effective variance weighted least squares algorithm incorporated in the CMB model considers both ambient and source data uncertainties. This fitting procedure gives less emphasis to those highly variable elements in the fitting process. If most elements in a source are highly variable, the source contribution estimate for that source is likely to have a high uncertainty.

APPENDIX B

Examples and Quick Guide for Identifying Problems and Corrective Action

There are four situations which, once they have been identified, can be addressed to improve the performance of the CMB. If they are not satisfactorily addressed, the model cannot be considered valid for a particular application or the source contribution estimates will have to be used with unacceptable uncertainties. The situations are: 1) incorrect ambient data; 2) incorrect source profiles; 3) missing source in the solution; 4) profile uncertainty/collinearity; and 5) noncontributing source in the solution.

The following provides a Quick Guide and examples to illustrate the process of using the model's statistics and diagnostics to help identify these situations. The indicators proposed in these examples are derived from a consensus of experienced model users and should be considered as "guides" -- not as "rules". They are not a "cure-all". They are only included to assist the discovery of errors in the input data which may cause the model to provide incorrect Source Contribution Estimates and they may be modified and improved in future revisions to this Protocol as a wide range of experience is gained in practical applications.

Example B is assumed to be the "correct" solution for the data set used in this example. Examples B.1 - B.5 show how the results would appear if the data were modified to include (one at a time) the situations identified above. Example B.6 shows a composite of several of these situations.

EXAMPLE B: THE "CORRECT" SOLUTION

SOURCE CONTRIBUTION ESTIMATES - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0
R SQUARE	.99	PERCENT MASS	101.8
CHI SQUARE	1.76	DF	9

SOURCE * TYPE	SCE(UG/M3)	STD ERR	TSTAT
3 UDUST	55.0169	1.8952	29.0298
4 AUTPB	2.8220	.5755	4.9033
5 RDOIL	.2834	.0811	3.4932
6 VBRN1	3.9257	1.5100	2.5998
11 ALPRO	2.1796	.9945	2.1917
13 FERMN	.1251	.0531	2.3535
17 SO4	1.3812	.5240	2.6357
18 NO3	1.3541	.3545	3.8194
19 OC	7.5042	4.5103	1.6638

MEASURED CONCENTRATION FINE/COARSE/TOTAL:
42.60000+- .600/ 73.30000+- 1.100/ 115.90000+- 1.253

UNCERTAINTY/SIMILARITY CLUSTERS — SUM OF COMB. SOURCES

SPECIES CONCENTRATIONS - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0		
R SQUARE	.99	PERCENT MASS	101.8		
CHI SQUARE	1.76	DF	9		

SPECIES-I-M		MEAS		CALC		RATIO C/M		RATIO R/U
1 TOT		73.30000+-	1.10000	74.59217+-	4.83016	1.02+- .07	TOT	.26
9 F	*	.03600+-	.01700	.09646+-	.05719	2.68+- 2.03	F	1.01
11 NA	*	.97000+-	.14000	1.05442+-	.05228	1.09+- .17	NA	.56
12 MG		.85000+-	.12000	.90611+-	.07983	1.07+- .18	MG	.39
13 AL	*	4.80000+-	.16000	4.40322+-	.40123	.92+- .09	AL	-.92
14 SI	*	14.50000+-	.50000	15.46888+-	.82643	1.07+- .07	SI	1.00
16 S		.56000+-	.14000	.52574+-	.04939	.94+- .25	S	-.23
17 CL	*	.33000+-	.05000	.32922+-	.06526	1.00+- .25	CL	-.01
19 K	*	.54000+-	.03000	.60618+-	.03237	1.12+- .09	K	1.50
20 CA	*	2.00000+-	.07000	1.75155+-	.12300	.88+- .07	CA	-1.76
22 TI	*	.40000+-	.02000	.55770+-	.09907	1.39+- .26	TI	1.56
23 V	*	.02600+-	.00200	.02551+-	.00349	.98+- .15	V	-.12
24 CR		.02600+-	.00200	.02529+-	.00937	.97+- .37	CR	-.07
25 MN	*	.08300+-	.00400	.08149+-	.00779	.98+- .11	MN	-.17
26 FE	*	3.50000+-	.13200	3.23851+-	.16125	.93+- .06	FE	-1.25
28 NI	*	.02200+-	.00300	.02248+-	.00403	1.02+- .23	NI	.09
29 CU		.05200+-	.00400	.02541+-	.00689	.49+- .14	CU	-3.3
30 ZN		.11000+-	.01000	.07245+-	.02069	.66+- .20	ZN	-1.63
35 BR	*	.22000+-	.02000	.15044+-	.04846	.68+- .23	BR	-1.33
82 PB	*	.62000+-	.07000	.76834+-	.11825	1.24+- .24	PB	1.08
91 OC	*	13.10000+-	4.30000	13.10000+-	1.04393	1.00+- .34	OC	.00
92 EC	*	1.70000+-	.90000	1.14293+-	.45013	.67+- .44	EC	-.55
93 SO4	*	1.70000+-	.50000	1.70000+-	.15074	1.00+- .31	SO4	.00
94 NO3	*	1.60000+-	.30000	1.60000+-	.17316	1.00+- .22	NO3	.00

ENTER COMMAND
SSCONT

$$\text{INDIVIDUAL RATIO} = \frac{\text{CALC SPECIES(PER SOURCE)}}{\text{MEAS SPECIES(ALL SOURCES)}}$$

SPEC↑SOURCE	3	4	5	6	11	13	17	18	19
1 TOTAL	.751	.038	.004	.054	.030	.002	.019	.018	.102
9 F	.122	.000	.004	.000	2.543	.010	.000	.000	.000
11 NA	.993	.000	.010	.026	.054	.004	.000	.000	.000
12 MG	.997	.000	.000	.000	.069	.000	.000	.000	.000
13 AL	.756	.006	.000	.012	.142	.000	.000	.000	.000
14 SI	1.062	.002	.000	.002	.000	.000	.000	.000	.000
16 S	.000	.020	.067	.034	.000	.004	.814	.000	.000
17 CL	.000	.257	.000	.660	.079	.002	.000	.000	.000
19 K	1.049	.004	.001	.044	.000	.024	.000	.000	.000
20 CA	.825	.018	.002	.021	.009	.001	.000	.000	.000
22 TI	1.389	.000	.001	.000	.004	.000	.000	.000	.000
23 V	.571	.000	.375	.000	.034	.001	.000	.000	.000
24 CR	.952	.000	.005	.000	.013	.002	.000	.000	.000
25 MN	.663	.000	.002	.057	.000	.261	.000	.000	.000
26 FE	.901	.017	.002	.002	.002	.001	.000	.000	.000
28 NI	.100	.023	.690	.000	.208	.000	.000	.000	.000
29 CU	.317	.040	.004	.068	.059	.001	.000	.000	.000
30 ZN	.550	.090	.010	.000	.002	.007	.000	.000	.000
35 BR	.020	.641	.000	.009	.012	.001	.000	.000	.000
82 PB	.328	.910	.001	.000	.000	.000	.000	.000	.000
91 OC	.140	.108	.002	.177	.000	.001	.000	.000	.573
92 EC	.502	.063	.005	.081	.021	.001	.000	.000	.000
93 SO4	.024	.022	.080	.037	.022	.003	.812	.000	.000
94 NO3	.007	.016	.001	.125	.000	.004	.000	.846	.000

PROBLEM B.1: INCORRECT AMBIENT DATA

INDICATION:

- o RATIO R/U << -2.0 for a species suggests either the ambient data are high or the profile data are low for the flagged species;
- o RATIO R/U >> 2.0 for a species would imply that the ambient data are low or the profile data for that species are high. If profile data are suspect, see Section 3.5.2 and problem B.2.

ACTION:

1. Review the uncertainty assigned to the species with the high residuals. Make any justifiable and appropriate changes and rerun the CMB. If this reduces the RATIO R/U, Step 2 is not necessary.

2. Delete the suspect species from the list of fitting species and rerun. If the SCE changes by one standard error, do not use this species in the fit until it has been remeasured.

EXAMPLE B.1: IDENTIFYING INCORRECT AMBIENT DATA

SUMMARY: A CHI SQUARE >> 4 and a % MASS of ~120% suggest that the fit is not satisfactory. The RATIO R/U for Si was << -2.0 implying high ambient data for that species (also, low profile data for the major source of that species or absence of a source of Si in the source list). The profile data were checked and were believed to be correct. The ambient data was then reviewed and it was found that the Si ambient data was erroneously entered as 29 ug/m3 for this example when the correct value was 14.5. Notice that the SCE's for UDUST and ALPRO are significantly changed with respect to Example B. The high ambient Si data raised the UDUST SCE to 69.6 ug/m^3 and the ALPRO SCE is lowered.

SOURCE CONTRIBUTION ESTIMATES - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0
R SQUARE	.93	PERCENT MASS	118.4
CHI SQUARE	11.54	DF	9

SOURCE * TYPE	SCE(UG/M3)	STD ERR	TSTAT
3 UDUST	69.5898	2.1824	31.8870
4 AUTPB	2.6964	.6017	4.4809
5 RDOIL	.2797	.0801	3.4912
6 VBRN1	3.4292	1.4281	2.4012
11 ALPRO	.5877	.8251	.7123
13 FERMN	.0257	.0580	.4433
17 SO4	1.4130	.5244	2.6947
18 NO3	1.3834	.3490	3.9634
19 OC	7.3824	4.5131	1.6358

MEASURED CONCENTRATION FINE/COARSE/TOTAL:
42.60000+- .600/ 73.30000+- 1.100/ 115.90000+- 1.253

UNCERTAINTY/SIMILARITY CLUSTERS SUM OF COMB. SOURCES

SPECIES CONCENTRATIONS - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0	
R SQUARE	.93	PERCENT MASS	118.4	
CHI SQUARE	11.54	DF	9	

SPECIES-I-M		MEAS		CALC		RATIO C/M		RATIO R/U
1 TOT		73.30000+-	1.10000	86.78726+-	4.94793	1.18+-	.07 TOT	2.66
9 F	*	.03600+-	.01700	.03048+-	.03639	.85+-	1.09 F	-.14
11 NA	*	.97000+-	.14000	1.26480+-	.06380	1.30+-	.20 NA	1.92
12 MG		.85000+-	.12000	1.08755+-	.09877	1.28+-	.21 MG	1.53
13 AL	*	4.80000+-	.16000	4.85758+-	.48899	1.01+-	.11 AL	.11
14 SI	*	29.00000+-	.50000	19.54129+-	1.04456	.67+-	.04 SI	-8.17
16 S		.56000+-	.14000	.53117+-	.05034	.95+-	.25 S	-.19
17 CL	*	.33000+-	.05000	.27837+-	.05884	.84+-	.22 CL	-.67
19 K	*	.54000+-	.03000	.74277+-	.04039	1.38+-	.11 K	4.03
20 CA	*	2.00000+-	.07000	2.16760+-	.15426	1.08+-	.09 CA	.99
22 TI	*	.40000+-	.02000	.70362+-	.12529	1.76+-	.33 TI	2.39
23 V	*	.02600+-	.00200	.02865+-	.00407	1.10+-	.18 V	.58
24 CR		.02600+-	.00200	.03155+-	.01183	1.21+-	.46 CR	.46
25 MN	*	.08300+-	.00400	.07828+-	.00871	.94+-	.11 MN	-.49
26 FE	*	3.50000+-	.13200	4.06171+-	.20300	1.16+-	.07 FE	2.32
28 NI	*	.02200+-	.00300	.01949+-	.00400	.89+-	.22 NI	-.50
29 CU		.05200+-	.00400	.02697+-	.00851	.52+-	.17 CU	-2.66
30 ZN		.11000+-	.01000	.08731+-	.02599	.79+-	.25 ZN	-.81
35 BR	*	.22000+-	.02000	.14299+-	.04630	.65+-	.22 BR	-1.53
82 PB	*	.62000+-	.07000	.79707+-	.13206	1.29+-	.26 PB	1.18
91 OC	*	13.10000+-	4.30000	13.10000+-	1.09582	1.00+-	.34 OC	.00
92 EC	*	1.70000+-	.90000	1.31958+-	.55518	.78+-	.52 EC	-.36
93 SO4	*	1.70000+-	.50000	1.70000+-	.15206	1.00+-	.31 SO4	.00
94 NO3	*	1.60000+-	.30000	1.60000+-	.16353	1.00+-	.21 NO3	.00

ENTER COMMAND
SSCONT

INDIVIDUAL RATIO = CALC SPECIES(PER SOURCE) / MEAS SPECIES(ALL SOURCES)

SPEC†SOURCE	3	4	5	6	11	13	17	18	19
1 TOTAL	.949	.037	.004	.047	.008	.000	.019	.019	.101
9 F	.155	.000	.004	.000	.686	.002	.000	.000	.000
11 NA	1.255	.000	.010	.023	.015	.001	.000	.000	.000
12 MG	1.261	.000	.000	.000	.019	.000	.000	.000	.000
13 AL	.957	.006	.000	.010	.038	.000	.000	.000	.000
14 SI	.672	.001	.000	.001	.000	.000	.000	.000	.000
16 S	.000	.019	.066	.029	.000	.001	.833	.000	.000
17 CL	.000	.245	.000	.577	.021	.000	.000	.000	.000
19 K	1.327	.004	.001	.038	.000	.005	.000	.000	.000
20 CA	1.044	.017	.002	.018	.002	.000	.000	.000	.000
22 TI	1.757	.000	.001	.000	.001	.000	.000	.000	.000
23 V	.723	.000	.370	.000	.009	.000	.000	.000	.000
24 CR	1.204	.000	.005	.000	.004	.000	.000	.000	.000
25 MN	.838	.000	.002	.050	.000	.054	.000	.000	.000
26 FE	1.139	.016	.002	.002	.001	.000	.000	.000	.000
28 NI	.127	.022	.681	.000	.056	.000	.000	.000	.000
29 CU	.401	.038	.004	.059	.016	.000	.000	.000	.000
30 ZN	.696	.086	.010	.000	.000	.001	.000	.000	.000
35 BR	.025	.613	.000	.008	.003	.000	.000	.000	.000
82 PB	.415	.870	.000	.000	.000	.000	.000	.000	.000
91 OC	.177	.103	.001	.154	.000	.000	.000	.000	.564
92 EC	.634	.060	.005	.071	.006	.000	.000	.000	.000
93 SO4	.030	.021	.079	.032	.006	.001	.831	.000	.000
94 NO3	.009	.015	.001	.109	.000	.001	.000	.865	.000

PROBLEM B.2: GROSS ERROR IN A SOURCE PROFILE

INDICATION:

- o SCE that is inconsistent with preliminary analyses or physical evidence;
- o one or more species has a "high (pos. or neg.)" residual which cannot be attributed to incorrect ambient data; further evidence if the SSCONT reveals that one source dominates that species.

ACTION: Review profile data for the suspect species carefully. Correct or remeasure profile if necessary.

EXAMPLE B.2: GROSS ERROR IN A SOURCE PROFILE

SUMMARY: The Ca RATIO R/U is << -2.0 implying high ambient data or low profile data for Ca and several other species. In SSCONT, the Ca is dominated by UDUST. Both the ambient and profile data were checked and no problem was found with Ca. The RATIO R/U for Fe was >> 2.0 implying low ambient data or high profile data for Fe. The ambient data were reviewed and Fe was found to be high in the UDUST profile (.114 instead of .057). This example points out that the diagnostics can only suggest possible sources of error. One real data error (in Fe) caused other species (Ca) to "appear" to be incorrect.

SOURCE CONTRIBUTION ESTIMATES - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0
R SQUARE	.94	PERCENT MASS	86.8
CHI SQUARE	11.51	DF	9

*	SOURCE TYPE	SCE(UG/M3)	STD ERR	TSTAT
3	UDUST	38.8685	1.2080	32.1758
4	AUTPB	2.7047	.5265	5.1368
5	RDOIL	.2403	.0878	2.7354
6	VBRN1	7.1187	2.1008	3.3886
11	ALPRO	5.7457	1.2366	4.6463
13	FERMN	.2239	.0638	3.5111
17	SO4	1.2995	.5278	2.4620
18	NO3	1.1902	.4004	2.9723
19	OC	6.2125	4.5956	1.3518

MEASURED CONCENTRATION FINE/COARSE/TOTAL:
42.60000+- .600/ 73.30000+- 1.100/ 115.90000+- 1.253

UNCERTAINTY/SIMILARITY CLUSTERS — SUM OF COMB. SOURCES

SPECIES CONCENTRATIONS - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0	
R SQUARE	.94	PERCENT MASS	86.8	
CHI SQUARE	11.51	DF	9	

SPECIES-I-M		MEAS		CALC		RATIO C/M		RATIO R/U
1 TOT		73.30000+-	1.10000	63.60400+-	4.72030	.87+-	.07 TOT	-2.00
9 F	*	.03600+-	.01700	.24521+-	.13036	6.81+-	4.84 F	1.59
11 NA	*	.97000+-	.14000	.87972+-	.05058	.91+-	.14 NA	-.61
12 MG		.85000+-	.12000	.75371+-	.06538	.89+-	.15 MG	-.70
13 AL	*	4.80000+-	.16000	4.49870+-	.39619	.94+-	.09 AL	-.71
14 SI	*	14.50000+-	.50000	10.97881+-	.58834	.76+-	.05 SI	-4.56
16 S		.56000+-	.14000	.50959+-	.05195	.91+-	.25 S	-.34
17 CL	*	.33000+-	.05000	.54612+-	.10871	1.65+-	.41 CL	1.81
19 K	*	.54000+-	.03000	.46919+-	.02488	.87+-	.07 K	-1.32
20 CA	*	2.00000+-	.07000	1.32928+-	.09279	.66+-	.05 CA	-5.77
22 TI	*	.40000+-	.02000	.39731+-	.07006	.99+-	.18 TI	-.04
23 V	*	.02600+-	.00200	.02111+-	.00273	.81+-	.12 V	-1.44
24 CR		.02600+-	.00200	.01862+-	.00669	.72+-	.26 CR	-1.06
25 MN	*	.08300+-	.00400	.08626+-	.00988	1.04+-	.13 MN	.31
26 FE	*	3.50000+-	.13200	4.53500+-	.11542	1.30+-	.06 FE	5.90
28 NI	*	.02200+-	.00300	.02699+-	.00466	1.23+-	.27 NI	.90
29 CU		.05200+-	.00400	.02835+-	.00598	.55+-	.12 CU	-3.29
30 ZN		.11000+-	.01000	.05500+-	.01482	.50+-	.14 ZN	-3.08
35 BR	*	.22000+-	.02000	.14940+-	.04763	.68+-	.23 BR	-1.37
82 PB	*	.62000+-	.07000	.68511+-	.10008	1.11+-	.20 PB	.53
91 OC	*	13.10000+-	4.30000	13.10000+-	1.06037	1.00+-	.34 OC	.00
92 EC	*	1.70000+-	.90000	1.05713+-	.39077	.62+-	.40 EC	-.66
93 SO4	*	1.70000+-	.50000	1.70000+-	.16139	1.00+-	.31 SO4	.00
94 NO3	*	1.60000+-	.30000	1.60000+-	.24324	1.00+-	.24 NO3	.00

ENTER COMMAND
SSCONT

$$\text{INDIVIDUAL RATIO} = \frac{\text{CALC SPECIES(PER SOURCE)}}{\text{MEAS SPECIES(ALL SOURCES)}}$$

SPEC↑SOURCE	3	4	5	6	11	13	17	18	19
1 TOTAL	.530	.037	.003	.097	.078	.003	.018	.016	.085
9 F	.086	.000	.004	.000	6.703	.018	.000	.000	.000
11 NA	.701	.000	.009	.048	.142	.007	.000	.000	.000
12 MG	.704	.000	.000	.000	.183	.000	.000	.000	.000
13 AL	.534	.006	.000	.021	.375	.000	.000	.000	.000
14 SI	.751	.002	.000	.004	.000	.000	.000	.000	.000
16 S	.000	.019	.057	.061	.000	.007	.766	.000	.000
17 CL	.000	.246	.000	1.197	.209	.003	.000	.000	.000
19 K	.741	.004	.001	.079	.000	.044	.000	.000	.000
20 CA	.583	.017	.002	.038	.023	.001	.000	.000	.000
22 TI	.981	.000	.001	.000	.011	.000	.000	.000	.000
23 V	.404	.000	.318	.000	.088	.002	.000	.000	.000
24 CR	.673	.000	.004	.000	.035	.004	.000	.000	.000
25 MN	.468	.000	.001	.103	.000	.467	.000	.000	.000
26 FE	1.266	.016	.002	.004	.006	.001	.000	.000	.000
28 NI	.071	.022	.585	.000	.548	.000	.000	.000	.000
29 CU	.224	.038	.003	.123	.155	.002	.000	.000	.000
30 ZN	.389	.086	.009	.000	.005	.012	.000	.000	.000
35 BR	.014	.615	.000	.017	.031	.002	.000	.000	.000
82 PB	.232	.872	.000	.000	.000	.000	.000	.000	.000
91 OC	.099	.103	.001	.321	.000	.002	.000	.000	.474
92 EC	.354	.060	.004	.147	.054	.002	.000	.000	.000
93 SO4	.017	.021	.068	.067	.057	.006	.764	.000	.000
94 NO3	.005	.015	.001	.227	.000	.008	.000	.744	.000

PROBLEM B.3: MISSING SOURCE

INDICATION:

- o High Chi-Square;
- o Low percent mass explained;
- o RATIO R/U << -2.0 (a "nigh negative" residual) for one or more species that are known to be present in the suspect source.

ACTION: Add sources to the fit and reevaluate.

EXAMPLE B.3: IDENTIFYING A MISSING SOURCE

SUMMARY: The Chi-Square is high and two species, Br and Pb have RATIO R/U << -2.0. Using the flowchart procedure in Figure 1, all data was rechecked and no problems were found. Thus, the source list was reviewed to check for missing sources. These low ratios for Br and Pb suggest a missing motor vehicle exhaust source. The percent mass explained is not low (a bit high, in fact), suggesting that a missing source may be a minor contributor. The absence of motor vehicle exhaust does not cause the other source contributions to be different from the "correct" ones because its source profile has few species in common with the other sources.

SOURCE CONTRIBUTION ESTIMATES - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0
R SQUARE	.91	PERCENT MASS	102.5
CHI SQUARE	12.16	DF	10

*	SOURCE TYPE	SCE(UG/M3)	STD ERR	TSTAT
3	UDUST	55.4251	1.9200	28.8676
5	RDOIL	.2620	.0831	3.1540
6	VBRN1	5.7480	1.7679	3.2513
11	ALPRO	3.0749	1.1644	2.6408
13	FERMN	.1072	.0610	1.7588
17	SO4	1.3842	.5259	2.6322
18	NO3	1.2880	.3729	3.4536
19	OC	7.8295	4.5613	1.7165

MEASURED CONCENTRATION FINE/COARSE/TOTAL:
42.60000+- .600/ 73.30000+- 1.100/ 115.90000+- 1.253

UNCERTAINTY/SIMILARITY CLUSTERS — SUM OF COMB. SOURCES

SPECIES CONCENTRATIONS - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0	
R SQUARE	.91	PERCENT MASS	102.5	
CHI SQUARE	12.16	DF	10	

SPECIES-I-M		MEAS		CALC		RATIO C/M		RATIO R/U
1 TOT		73.30000+-	1.10000	75.11888+-	4.86535	1.02+- .07	TOT	.36
9 F	*	.03600+-	.01700	.13403+-	.08205	3.72+- 2.88	F	1.17
11 NA	*	.97000+-	.14000	1.09359+-	.05449	1.13+- .17	NA	.82
12 MG		.85000+-	.12000	.93657+-	.07992	1.10+- .18	MG	.60
13 AL	*	4.80000+-	.16000	4.70535+-	.41893	.98+- .09	AL	-.21
14 SI	*	14.50000+-	.50000	15.57675+-	.83378	1.07+- .07	SI	1.11
16 S		.56000+-	.14000	.52105+-	.05130	.93+- .25	S	-.26
17 CL	*	.33000+-	.05000	.35636+-	.08534	1.08+- .31	CL	.27
19 K	*	.54000+-	.03000	.61736+-	.03257	1.14+- .09	K	1.75
20 CA	*	2.00000+-	.07000	1.75470+-	.12433	.88+- .07	CA	-1.72
22 TI	*	.40000+-	.02000	.56247+-	.09977	1.41+- .26	TI	1.60
23 V	*	.02600+-	.00200	.02523+-	.00341	.97+- .15	V	-.19
24 CR		.02600+-	.00200	.02560+-	.00944	.98+- .37	CR	-.04
25 MN	*	.08300+-	.00400	.08100+-	.00929	.98+- .12	MN	-.20
26 FE	*	3.50000+-	.13200	3.20850+-	.16094	.92+- .06	FE	-1.40
28 NI	*	.02200+-	.00300	.02271+-	.00403	1.03+- .23	NI	.14
29 CU		.05200+-	.00400	.02634+-	.00714	.51+- .14	CU	-3.13
30 ZN		.11000+-	.01000	.06291+-	.02051	.57+- .19	ZN	-2.06
35 BR	*	.22000+-	.02000	.01138+-	.00832	.05+- .04	BR	-9.63
82 PB	*	.62000+-	.07000	.20541+-	.08319	.33+- .14	PB	-3.81
91 OC	*	13.10000+-	4.30000	13.10000+-	1.11467	1.00+- .34	OC	.00
92 EC	*	1.70000+-	.90000	1.11920+-	.47212	.66+- .45	EC	-.57
93 SO4	*	1.70000+-	.50000	1.70000+-	.15639	1.00+- .31	SO4	.00
94 NO3	*	1.60000+-	.30000	1.60000+-	.20264	1.00+- .23	NO3	.00

ENTER COMMAND
SSCONT

INDIVIDUAL RATIO = CALC SPECIES(PER SOURCE) / MEAS SPECIES(ALL SOURCES)

SPEC↑SOURCE	3	5	6	11	13	17	18	19
1 TOTAL	.756	.004	.078	.042	.001	.019	.018	.107
9 F	.123	.004	.000	3.587	.009	.000	.000	.000
11 NA	1.000	.009	.039	.076	.003	.000	.000	.000
12 MG	1.004	.000	.000	.098	.000	.000	.000	.000
13 AL	.762	.000	.017	.201	.000	.000	.000	.000
14 SI	1.070	.000	.004	.000	.000	.000	.000	.000
16 S	.000	.062	.049	.000	.003	.816	.000	.000
17 CL	.000	.000	.967	.112	.001	.000	.000	.000
19 K	1.057	.001	.064	.000	.021	.000	.000	.000
20 CA	.831	.002	.031	.012	.001	.000	.000	.000
22 TI	1.399	.001	.000	.006	.000	.000	.000	.000
23 V	.576	.347	.000	.047	.001	.000	.000	.000
24 CR	.959	.005	.000	.019	.002	.000	.000	.000
25 MN	.668	.001	.083	.000	.224	.000	.000	.000
26 FE	.907	.002	.003	.003	.001	.000	.000	.000
28 NI	.101	.638	.000	.294	.000	.000	.000	.000
29 CU	.320	.004	.099	.083	.001	.000	.000	.000
30 ZN	.554	.010	.000	.003	.006	.000	.000	.000
35 BR	.020	.000	.014	.017	.001	.000	.000	.000
82 PB	.331	.000	.000	.000	.000	.000	.000	.000
91 OC	.141	.001	.259	.000	.001	.000	.000	.598
92 EC	.505	.005	.118	.029	.001	.000	.000	.000
93 SO4	.024	.074	.054	.031	.003	.814	.000	.000
94 NO3	.007	.001	.183	.000	.004	.000	.805	.000

PROBLEM B.4: COLLINEARITY

INDICATION:

- o Two or more sources listed in a U/S cluster
- o T-statistic < 2.0 for one or more sources in that cluster-- if the T-STAT becomes > 2.0 when species uncertainties for profile for that source is arbitrarily reduced to a potentially achievable level, this indicates that the uncertainty in the source profile is at least partially responsible for the "apparent" collinearity.

ACTION: Remedies for unacceptably high uncertainties due to collinearity can take four forms ranked from most to least desirable. There is an additional remedy if the collinearity is associated with high uncertainties as discussed above.

EXAMPLE B.4: COLLINEARITY

SUMMARY: VBRN2 is a source that was added to the source list to introduce collinearity. In this case, VBRN2 was actually not contributing but it illustrates the possibility that two potentially legitimate sources may have similar signatures causing collinearity. Following the Figure 1 flowchart, all source profile and species data have been checked and no missing sources are apparent at this time. The only indications of a problem are the negative SCE for VBRN2 and the fact that VBRN1, VBRN2 and OC sources were all flagged in U/S clusters, indicating collinearity. The SSCONT screen suggests that removing Cl from the fit might reduce the collinearity. Other approaches would include deleting one or two of the sources from the fit. The uncertainty/similarity display suggests that the combined contribution of sources 6 & 19 is 13.9 +- 5.1 and the sum of 6,7, and 19 is only 12.1 +- 4.7. This suggests that VBRN2 (source #7) may be an insignificant contributing source and could be deleted. Other options for dealing with collinearity are discussed in Section 3.5.4. Only VBRN1's SCE is significantly affected by the addition of VBRN2, but its STD. ERROR nearly triples with respect to the "true" solution (Example B). Notice that the "true" VBRN1 still falls within one uncertainty interval of the VRRN1 contribution in this example. This example graphically illustrates the effects of collinearity on source contribution estimates and their uncertainties.

SOURCE CONTRIBUTION ESTIMATES - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0
R SQUARE	.99	PERCENT MASS	103.9
CHI SQUARE	1.45	DF	8

SOURCE *	TYPE	SCE(UG/M3)	STD ERR	TSTAT
3	UDUST	55.9124	2.1142	26.4464
4	AUTPB	2.7711	.5742	4.8263
5	RDOIL	.2748	.0838	3.2791
6	VBRN1	8.0754	4.3667	1.8493
7	VBRN2	-1.7924	1.6701	-1.0733
11	ALPRO	2.3581	1.2212	1.9310
13	FERMN	.1061	.0758	1.4003
17	SO4	1.4063	.5395	2.6067
18	NO3	1.1798	.4333	2.7226
19	OC	5.8962	4.9430	1.1928

MEASURED CONCENTRATION FINE/COARSE/TOTAL:
42.60000+- .600/ 73.30000+- 1.100/ 115.90000+- 1.253

UNCERTAINTY/SIMILARITY CLUSTERS	SUM OF COMB. SOURCES
6 19	13.972+- 5.144
6 7 19	12.179+- 4.700
7 11	.566+- 2.172
6 7 11	8.641+- 3.305

SPECIES CONCENTRATIONS - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0
R SQUARE	.99	PERCENT MASS	103.9
CHI SQUARE	1.45	DF	8

SPECIES-I-M		MEAS		CALC		RATIO C/M		RATIO R/U
1 TOT		73.30000+-	1.10000	76.18761+-	5.01175	1.04+- .07	TOT	.56
9 F	*	.03600+-	.01700	.09823+-	.09247	2.73+- 2.87	F	.66
11 NA	*	.97000+-	.14000	1.09454+-	.05457	1.13+- .17	NA	.83
12 MG		.85000+-	.12000	.92472+-	.08303	1.09+- .18	MG	.51
13 AL	*	4.80000+-	.16000	4.56913+-	.41389	.95+- .09	AL	-.52
14 SI	*	14.50000+-	.50000	15.74725+-	.84354	1.09+- .07	SI	1.27
16 S		.56000+-	.14000	.52360+-	.06079	.93+- .26	S	-.24
17 CL	*	.33000+-	.05000	.38261+-	.15282	1.16+- .50	CL	.33
19 K	*	.54000+-	.03000	.52175+-	.08708	.97+- .17	K	-.20
20 CA	*	2.00000+-	.07000	1.80675+-	.12900	.90+- .07	CA	-1.32
22 TI	*	.40000+-	.02000	.56560+-	.10069	1.41+- .26	TI	1.61
23 V	*	.02600+-	.00200	.02552+-	.00349	.98+- .15	V	-.12

24 CR		.02600+-	.00200	.02550+-	.00952	.98+-	.37 CR	-.05
25 MN	*	.08300+-	.00400	.08324+-	.01157	1.00+-	.15 MN	.02
26 FE	*	3.50000+-	.13200	3.29570+-	.16395	.94+-	.06 FE	-.97
28 NI	*	.02200+-	.00300	.02242+-	.00399	1.02+-	.23 NI	.08
29 CU		.05200+-	.00400	.02864+-	.00763	.55+-	.15 CU	-2.71
30 ZN		.11000+-	.01000	.07313+-	.02101	.66+-	.20 ZN	-1.58
35 BR	*	.22000+-	.02000	.14953+-	.04775	.68+-	.23 BR	-1.36
82 PB	*	.62000+-	.07000	.76144+-	.11811	1.23+-	.24 PB	1.03
91 OC	*	13.10000+-	4.30000	13.10000+-	1.34876	1.00+-	.34 OC	.00
92 EC	*	1.70000+-	.90000	1.22355+-	.53045	.72+-	.49 EC	-.46
93 SO4	*	1.70000+-	.50000	1.70000+-	.19156	1.00+-	.31 SO4	.00
94 NO3	*	1.60000+-	.30000	1.60000+-	.24294	1.00+-	.24 NO3	.00

ENTER COMMAND
SSCONT

$$\text{INDIVIDUAL RATIO} = \frac{\text{CALC SPECIES(PER SOURCE)}}{\text{MEAS SPECIES(ALL SOURCES)}}$$

SPEC↑SOURCE	3	4	5	6	7	11	13	17	18	19
1 TOTAL	.763	.038	.004	.110	-.024	.032	.001	.019	.016	.080
9 F	.124	.000	.004	.000	-.159	2.751	.009	.000	.000	.000
11 NA	1.009	.000	.010	.054	-.006	.058	.003	.000	.000	.000
12 MG	1.013	.000	.000	.000	.000	.075	.000	.000	.000	.000
13 AL	.769	.006	.000	.024	-.002	.154	.000	.000	.000	.000
14 SI	1.080	.002	.000	.005	-.001	.000	.000	.000	.000	.000
16 S	.000	.020	.065	.069	-.051	.000	.003	.829	.000	.000
17 CL	.000	.252	.000	1.358	-.538	.086	.001	.000	.000	.000
19 K	1.066	.004	.001	.090	-.216	.000	.021	.000	.000	.000
20 CA	.839	.017	.002	.043	-.008	.010	.001	.000	.000	.000
22 TI	1.412	.000	.001	.000	-.003	.004	.000	.000	.000	.000
23 V	.581	.000	.364	.000	.000	.036	.001	.000	.000	.000
24 CR	.968	.000	.005	.000	-.008	.015	.002	.000	.000	.000
25 MN	.674	.000	.002	.117	-.010	.000	.221	.000	.000	.000
26 FE	.915	.017	.002	.004	.000	.003	.001	.000	.000	.000
28 NI	.102	.023	.669	.000	.000	.225	.000	.000	.000	.000
29 CU	.323	.039	.004	.140	-.019	.063	.001	.000	.000	.000
30 ZN	.559	.088	.010	.000	.000	.002	.006	.000	.000	.000
35 BR	.020	.630	.000	.019	-.004	.013	.001	.000	.000	.000
82 PB	.334	.894	.000	.000	.000	.000	.000	.000	.000	.000
91 OC	.143	.106	.001	.364	-.064	.000	.001	.000	.000	.450
92 EC	.510	.062	.005	.166	-.046	.022	.001	.000	.000	.000
93 SO4	.024	.021	.078	.076	-.053	.024	.003	.827	.000	.000
94 NO3	.007	.016	.001	.257	-.022	.000	.004	.000	.737	.000

PROBLEM B.5: NONCONTRIBUTING SOURCE IN FIT CAUSING COLLINEARITY

INDICATION:

- o T-STAT between -2.0 and 2.0
- o RATIO R/U >> 2.0 ("high positive" residual) for a species which is attributed to the suspect source by the SSCONT diagnostic
- o Negative SCE
- o Physical basis for the source's contribution is weak.
- o SCE's, statistics and diagnostics do not change if the suspect source is deleted

ACTION: Delete source from fit

EXAMPLE B.5: REMOVING NONCONTRIBUTING SOURCES FROM THE FIT

SUMMARY: All source profile and species data were reviewed and found to be correct. At this point, no justification can be made tnat a missing source must be added. The SCE for the GLASS source is negative and has the lowest T-STAT of any SCE. The organic carbon (OC) source has a T-STAT < 2.0 and the magnitude of the STDERR is high (4.5), so it may well be a minor or noncontributor. The uncertainty/similarity display shows that there is significant collinearity involving the GLASS source. There is no "high positive" RATIO R/U to indicate that GLASS is a noncontributor. In this case, the fact that deletion of the GLASS source did not affect the SCE's, statistics and diagnostics coupled with its negative SCE and low T-STAT were used to conclude that the GLASS was noncontributing.

SOURCE CONTRIBUTION ESTIMATES - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0
R SQUARE	.99	PERCENT MASS	102.2
CHI SQUARE	1.73	DF	8

SOURCE * TYPE	SCE(UG/M3)	STD ERR	TSTAT
3 UDUST	55.9343	2.0525	27.2514
4 AUTPB	2.8157	.5775	4.8755
5 RDOIL	.2801	.0814	3.4399
6 VBRN1	4.0938	1.5473	2.6458
11 ALPRO	2.2675	1.0213	2.2202
13 FERMN	.1217	.0542	2.2429
15 GLASS	-2.0215	1.6002	-1.2633
17 SO4	2.6921	1.2022	2.2394
18 NO3	1.3582	.3565	3.8095
19 OC	7.3781	4.5155	1.6340

MEASURED CONCENTRATION FINE/COARSE/TOTAL:
42.60000+- .600/ 73.30000+- 1.100/ 115.90000+- 1.253

UNCERTAINTY/SIMILARITY CLUSTERS	SUM OF COMB. SOURCES	
3 15 17	56.605+-	2.026
3 6 15 17	60.699+-	2.431
3 6 11 15	60.274+-	2.384
15 17	.671+-	.327

SPECIES CONCENTRATIONS - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0	
R SQUARE	.99	PERCENT MASS	102.2	
CHI SQUARE	1.73	DF	8	

SPECIES-I-M		MEAS		CALC		RATIO C/M		RATIO R/U	
1 TOT		73.30000+-	1.10000	74.91993+-	4.85608	1.02+-	.07	TOT	.33
9 F	*	.03600+-	.01700	.09912+-	.05956	2.75+-	2.10	F	1.02
11 NA	*	.97000+-	.14000	.88748+-	.05694	.91+-	.14	NA	-.55
12 MG		.85000+-	.12000	.92261+-	.08360	1.09+-	.18	MG	.50
13 AL	*	4.80000+-	.16000	4.48975+-	.40866	.94+-	.09	AL	-.71
14 SI	*	14.50000+-	.50000	15.72217+-	.84027	1.08+-	.07	SI	1.25
16 S		.56000+-	.14000	.73020+-	.10215	1.30+-	.37	S	.98
17 CL	*	.33000+-	.05000	.33818+-	.06746	1.02+-	.26	CL	.10
19 K	*	.54000+-	.03000	.58858+-	.03336	1.09+-	.09	K	1.08
20 CA	*	2.00000+-	.07000	1.77534+-	.12508	.89+-	.07	CA	-1.57
22 TI	*	.40000+-	.02000	.56702+-	.10073	1.42+-	.26	TI	1.63
23 V	*	.02600+-	.00200	.02558+-	.00351	.98+-	.15	V	-.10
24 CR		.02600+-	.00200	.02188+-	.00955	.84+-	.37	CR	-.42
25 MN	*	.08300+-	.00400	.08198+-	.00799	.99+-	.11	MN	-.11
26 FE	*	3.50000+-	.13200	3.29103+-	.16388	.94+-	.06	FE	-.99
28 NI	*	.02200+-	.00300	.02252+-	.00403	1.02+-	.23	NI	.10
29 CU		.05200+-	.00400	.02585+-	.00702	.50+-	.14	CU	-3.24
30 ZN		.11000+-	.01000	.07319+-	.02103	.67+-	.20	ZN	-1.58
35 BR	*	.22000+-	.02000	.15025+-	.04838	.68+-	.23	BR	-1.33
82 PB	*	.62000+-	.07000	.76278+-	.11909	1.23+-	.24	PB	1.03
91 OC	*	13.10000+-	4.30000	13.10000+-	1.04598	1.00+-	.34	OC	.00
92 EC	*	1.70000+-	.90000	1.16405+-	.45855	.68+-	.45	EC	-.53
93 SO4	*	1.70000+-	.50000	1.70000+-	.34105	1.00+-	.36	SO4	.00
94 NO3	*	1.60000+-	.30000	1.60000+-	.17648	1.00+-	.22	NO3	.00

ENTER COMMAND
SSCONT

$$\text{INDIVIDUAL RATIO} = \frac{\text{CALC SPECIES(PER SOURCE)}}{\text{MEAS SPECIES(ALL SOURCES)}}$$

SPEC↑SOURCE	3	4	5	6	11	13	15	17	18	19
1 TOTAL	.763	.038	.004	.056	.031	.002	-.028	.037	.019	.101
9 F	.124	.000	.004	.000	2.645	.010	-.030	.000	.000	.000
11 NA	1.009	.000	.010	.027	.056	.004	-.192	.000	.000	.000
12 MG	1.013	.000	.000	.000	.072	.000	.000	.000	.000	.000
13 AL	.769	.006	.000	.012	.148	.000	-.001	.000	.000	.000
14 SI	1.080	.002	.000	.003	.000	.000	.000	.000	.000	.000
16 S	.000	.020	.067	.035	.000	.004	-.408	1.586	.000	.000
17 CL	.000	.256	.000	.689	.082	.002	-.004	.000	.000	.000
19 K	1.067	.004	.001	.045	.000	.024	-.051	.000	.000	.000
20 CA	.839	.018	.002	.022	.009	.001	-.003	.000	.000	.000
22 TI	1.412	.000	.001	.000	.004	.000	.000	.000	.000	.000
23 V	.581	.000	.371	.000	.035	.001	-.004	.000	.000	.000
24 CR	.968	.000	.005	.000	.014	.002	-.148	.000	.000	.000
25 MN	.674	.000	.002	.059	.000	.254	-.001	.000	.000	.000
26 FE	.916	.017	.002	.002	.002	.001	.000	.000	.000	.000
28 NI	.102	.023	.682	.000	.216	.000	.000	.000	.000	.000
29 CU	.323	.040	.004	.071	.061	.001	-.002	.000	.000	.000
30 ZN	.559	.090	.010	.000	.002	.006	-.002	.000	.000	.000
35 BR	.020	.640	.000	.010	.012	.001	-.001	.000	.000	.000
82 PB	.334	.908	.000	.000	.000	.000	-.012	.000	.000	.000
91 OC	.143	.107	.001	.184	.000	.001	.000	.000	.000	.563
92 EC	.510	.063	.005	.084	.021	.001	.000	.000	.000	.000
93 SO4	.024	.022	.079	.039	.023	.003	-.773	1.584	.000	.000
94 NO3	.007	.016	.001	.130	.000	.004	-.008	.000	.849	.000

EXAMPLE B.6: COMPOSITE - PROBLEMS B.1, B.2, B.3, B.4, B.5

SOURCE CONTRIBUTION ESTIMATES - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0
R SQUARE	.73	PERCENT MASS	106.8
CHI SQUARE	44.78	DF	8

SOURCE *	TYPE	SCE(UG/M3)	STD ERR	TSTAT
3	UDUST	41.9295	1.5526	27.0065
5	RDOIL	.2729	.1488	1.8343
6	VBRN1	63.8163	18.1976	3.5069
7	VBRN2	-6.8923	5.5143	-1.2499
11	ALPRO	3.3383	2.1902	1.5242
13	FERMN	-.2333	.4555	-.5122
15	GLASS	-1.6067	2.1410	-.7504
17	SO4	1.8586	1.6012	1.1608
18	NO3	-1.5037	1.8503	-.8127
19	OC	-22.7108	12.6036	-1.8019

INDICATIONS:

o Chi-Square is >> 4, indicating that the fit of the data is not satisfactory.

o RATIO R/U << -2.0 for silicon (Si) suggesting that either: 1) the ambient Si data are high; 2) profile(s) containing Si are low, or 3) there is a missing source which is dominated by Si.

o RATIO R/U << -2.0 for Pb and Br suggesting bad ambient data, bad profile data, or a missing source, as above.

o RATIO R/U >> 2.0 for Fe suggesting that either: 1) ambient Fe data are low or profiles containing Fe are high.

o Minor R/U problems for Cl, Zn.

o Evidence of collinearity involving sources 6, 7, and 19.

o Evidence of an undetermined problem with source 15.

ACTION: Using the flowchart in Figure 1, the data problems would be first addressed: Si would be resolved as in problem B.1, and Fe would be resolved as in problem B.2. Then, the Pb-Br problem is resolved as a missing source as in problem B.3. Following this, the collinearity problems would be addressed as in B.4 and B.5.

MEASURED CONCENTRATION FINE/COARSE/TOTAL:
42.60000+- .600/ 73.30000+- 1.100/ 115.90000+- 1.253

UNCERTAINTY/SIMILARITY CLUSTERS	SUM OF COMB. SOURCES	
6 19		
6 7 19	41.106+-	12.539
6 7	34.213+-	10.441
11 15 17	56.924+-	15.480
3 11 15 17	3.590+-	2.510
3 11	45.520+-	2.895
15 17	45.268+-	2.595
	.252+-	1.226

SPECIES CONCENTRATIONS - SITE:PACS2 DATE:0124 78 SIZE:COARSE

SAMPLE DURATION	24	START HOUR	0	
R SQUARE	.73	PERCENT MASS	106.8	
CHI SQUARE	44.78	DF	8	

SPECIES-I-M		MEAS		CALC		RATIO C/M		RATIO R/U
1 TOT		73.30000+-	1.10000	78.26896+-	9.61944	1.07+- .13	TOT	.51
9 F	*	.03600+-	.01700	.12011+-	.64140	3.34+-17.89	F	.13
11 NA	*	.97000+-	.14000	1.06045+-	.10329	1.09+- .19	NA	.52
12 MG		.85000+-	.12000	.73585+-	.09458	.87+- .17	MG	-.75
13 AL	*	4.80000+-	.16000	4.69708+-	.62825	.98+- .13	AL	-.16
14 SI	*	29.00000+-	.50000	12.27399+-	.94406	.42+- .03	SI	-15.66
16 S		.56000+-	.14000	.66017+-	.22953	1.18+- .50	S	.37
17 CL	*	.33000+-	.05000	2.89758+-	1.00184	8.78+- 3.31	CL	2.56
19 K	*	.54000+-	.03000	.32103+-	.31140	.59+- .58	K	-.70
20 CA	*	2.00000+-	.07000	1.90081+-	.26835	.95+- .14	CA	-.36
22 TI	*	.40000+-	.02000	.42139+-	.07564	1.05+- .20	TI	.27
23 V	*	.02600+-	.00200	.02191+-	.00296	.84+- .13	V	-1.14
24 CR		.02600+-	.00200	.01555+-	.00720	.60+- .28	CR	-1.40
25 MN	*	.08300+-	.00400	.07500+-	.07681	.90+- .93	MN	-.10
26 FE	*	3.50000+-	.13200	4.91307+-	.14060	1.40+- .07	FE	7.33
28 NI	*	.02200+-	.00300	.02331+-	.00407	1.06+- .23	NI	.26
29 CU		.052005	.00400	.07101+-	.02638	1.37+- .52	CU	.71
30 ZN		.11000+-	.01000	.04598+-	.01553	.42+- .15	ZN	-3.46
35 BR	*	.22000+-	.02000	.03764+-	.02928	.17+- .13	BR	-5.14
82 PB	*	.62000+-	.07000	.14923+-	.06299	.24+- .11	PB	-5.00
91 OC	*	13.10000+-	4.30000	13.10001+-	7.24982	1.00+- .64	OC	.00
92 EC	*	1.70000+-	.90000	2.63859+-	2.14457	1.55+- 1.51	EC	.40
93 SO4	*	1.70000+-	.50000	1.70000+-	.71868	1.00+- .52	SO4	.00
94 NO3	*	1.60000+-	.30000	1.60000+-	1.61313	1.00+- 1.03	NO3	.00

ENTER COMMAND
SSCONT

$$\text{INDIVIDUAL RATIO} = \frac{\text{CALC SPECIES(PER SOURCE)}}{\text{MEAS SPECIES(ALL SOURCES)}}$$

SPEC↑SOURCE	3	5	6	7	11	13	15	17	18	19
1 TOTAL	.572	.004	.871	-.094	.046	-.003	-.022	.025	-.021	-.310
9 F	.093	.004	.000	-.613	3.895	-.019	-.024	.000	.000	.000
11 NA	.756	.010	.428	-.023	.083	-.007	-.152	.000	.000	.000
12 MG	.760	.000	.000	.000	.106	.000	.000	.000	.000	.000
13 AL	.577	.000	.191	-.006	.218	.000	-.001	.000	.000	.000
14 SI	.405	.000	.020	-.001	.000	.000	.000	.000	.000	.000
16 S	.000	.065	.547	-.197	.000	-.007	-.324	1.095	.000	.000
17 CL	.000	.000	10.733	-2.068	.121	-.003	-.003	.000	.000	.000
19 K	.800	.001	.709	-.830	.000	-.045	-.041	.000	.000	.000
20 CA	.629	.002	.341	-.032	.014	-.002	-.002	.000	.000	.000
22 TI	1.059	.001	.000	-.012	.006	.000	.000	.000	.000	.000
23 V	.435	.361	.000	.000	.051	-.002	-.003	.000	.000	.000
24 CR	.726	.005	.000	-.032	021	-.004	-.117	.000	.000	.000
25 MN	.505	.002	.923	-.039	.000	-.486	.000	.000	.000	.000
26 FE	1.366	.002	.035	-.001	.004	-.001	.000	.000	.000	.000
28 NI	.076	.665	.000	.000	.319	.000	.000	.000	.000	.000
29 CU	.242	.004	1.105	-.072	.090	-.002	-.001	.000	.000	.000
30 ZN	.419	.010	.000	.000	.003	-.012	-.002	.000	.000	.000
35 BR	.015	.000	.154	-.014	.018	-.002	.000	.000	.000	.000
82 PB	.250	.000	.000	.000	.000	.000	-.010	.000	.000	.000
91 OC	.107	.001	2.874	-.247	.000	-.002	.000	.000	.000	-1.734
92 EC	.382	.005	1.314	-.178	.031	-.002	.000	.000	.000	.000
93 SO4	.018	.077	.601	-.203	.033	-.006	-.614	1.093	.000	.000
94 NO3	.005	.001	2.034	-.086	.000	-.008	-.006	.000	-.940	.000

APPENDIX C

CMB Model Assumptions

C.1 Constant Composition of Source Emissions

The chemical profile of particulate emissions must remain constant during transport between source and receptor. There are two types of changes which may occur. First, if the chemical composition of emitted particles differs substantially with particle size, deposition (settling or removal) of the larger particles could alter the source profile as perceived at the receptor. For instance, a particular element might be much more prevalent in the fine particles (< 2.5um) than in the coarse ones (> 2.5um) because of the nature of the source's operation. If the source profiles were available for the fine and coarse particles separately, there would be little cause for concern; however, if only a "total" profile (fine and coarse together) were available, deposition, which primarily affects the coarse particles, would effectively alter the profile between source and receptor.

Remedial actions could include re-measurement of the profile in the separate size fractions, or deletion from the profile of species that are affected by the selective deposition of large particles.

Second, gases that react and transform into particles during transport between source and receptor are often not represented in the source profiles. One such gas that requires careful consideration is SO_2, because of its potential for conversion to sulfate. This is discussed further in Section 3.2.1. It may be possible to construct a model that fits on both particulate and gaseous sulfur (e.g., Chow, 1985, Sheff and Wadden, 1985).

C.2 Proper Source Identification and Characterization

Incorrect source identification takes one of three forms:

- o Contributing source types are missing.
- o Non-contributing source types have been included.
- o Several sets of source profiles provide equally precise but significantly different source composition estimates.

C.2.1 Missing Source Types

The most common source types are those of secondary compounds of sulfate, nitrate, and organic carbon. These substances are emitted as gases which later turn into particles which are measured at the receptor. Most source profiles apply only to primary species. Several methods have been proposed for dealing with secondary species, none of which are completely satisfactory or proven.

The most common, and easiest method, is the single constituent source type proposed by Watson (1979). This source profile contains only one entry for the constituent (i.e., sulfate, nitrate, or organic carbon) which is the result of secondary aerosol formation. This "source" will account for the secondary contributors to these species, but it cannot identify the individual source types contributing the secondary components. The sulfate, nitrate and organic carbon which is primary (i.e., emitted as that species as opposed to being formed in the atmosphere) will be accounted for by the source profile measured at the emission point. These individual contributors can be identified by using the PSCONT command.

Another method, proposed by Chow (1985) and Scheff and Wadden (1985) is the combination of gaseous and particulate compounds containing the secondary species. Very few source profiles and ambient data which are currently available contain adequate measurements of gaseous and particulate phases, and this method would only be applicable in a Level III study (e.g., Blumenthal and Watson, 1986; Chow et al., 1986; Liu et al., 1986).

A final method for dealing with the missing secondary source types involves modifying the source profiles by a fractionation coefficient calculated from transport times and deposition and transformation rates. These fractionation factors might also be determined from emissions-aging experiments in enclosed environments such as smog chambers (Liu et al., 1986). Research on such fractionation coefficients is in progress (e.g., Friedlander, 1981; Stafford and Liljestrand, 1984; Lewis and Stevens, 1985; Chow, 1985) which might be applicable in certain situations. These fractionation coefficients are unvalidated, however, and may add more uncertainty than they eliminate.

C.2.2 Non (or Low) - Contributing Source Types

Low contributing sources generally do not affect the SCE for other sources if their profiles are not similar to them. Therefore, they need not be eliminated from the fit unless they are shown to be truly very minor contributors (see 3.5.3.2) and their profiles are collinear with other sources.

C.2.3 Conflicting Results

In the process of applying the CMB to different sets of chemical species and source types for a given receptor sample, it is common to find more than one "fit" which possesses a low Chi-Square, has a high R-Square, and which accounts for most of the chemical concentrations measured in the receptor samples. This was demonstrated in an example by Watson (1979) from which he concluded that "the receptor model tells what could be the contributors, not necessarily what are the contributors." It is important to try all suspected source types in the fit to help identify if this situation exists. In such case, additional physical information is then needed to determine which of them is likely to be true. This may be derived from wind direction analysis, microscopic examination of the ambient sample, operating schedules, dispersion modeling, etc.

C.3 Linear Independence of Profiles

The source profiles must be linearly independent or as nearly so as possible. Collinear, or chemically similar source profiles can interact to introduce large uncertainties into the SCE. There is always some similarity among source profiles, because there are a finite number of elements or species that are readily measurable, and a much larger number of sources. High source profile uncertainty can exacerbate collinearity caused by similar profiles or can result in large standard errors.

As explained in Section 3.3.4, inherent uncertainty is a measure of this collinearity and high uncertainty in source profiles and, thus, the capability of the model to distinguish among the sources included in the model run. The sources within each U/S cluster are those which are largely responsible for the inherent uncertainty of the cluster through their interaction with each other and the uncertainty associated with their individual species (after Henry, 1982).

Innerent uncertainty is not a concept which the CMB user must understand in detail. The user must, however, appreciate that a high inherent uncertainty is caused by groups of sources that have somewhat similar profiles and/or by profile uncertainties which are sufficiently large such that the model cannot distinguish among them with acceptable standard errors. This ultimately results in higher standard errors for the SCE's. If source profiles were known exactly (profile uncertainty equal to zero), the model would tolerate a fairly high collinearity and still provide stable and acceptable SCE's. However, source profiles generally are not known exactly and this imprecision can both reduce the amount of collinearity the model can tolerate and increase the standard error of the SCE's for the sources in the uncertainty cluster.

There are two possible means to reduce the standard errors of their SCE's, which are listed within U/S clusters. The first is to measure additional species (at both the source and the receptor) which will allow these sources to be differentiated from each other by the CMB. The second is to reduce the uncertainties in the source profiles in the cluster by making more precise source profile measurements.

In order to identify the appropriate means to improve the source contribution estimate (i.e., reduce uncertainty in the profile or measure additional species to reduce collinearity) one must determine whether the high inherent uncertainty of the cluster must be attributed to collinearity or to source profile uncertainties. A simple test is proposed in section 3.5.4.

C.4 Additional CMB Assumptions

In addition to the above four assumptions, the number of source categories must be less than the number of species included in the fit. Also, measurement errors must be random, uncorrelated and normally distributed.